CONTEXTUALIZATION IN LARGE SCALE SOCIAL NETWORKS

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Rizwana Irfan

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Electrical and Computer Engineering

December 2014

Fargo, North Dakota

North Dakota State University

Graduate School

Title

CONTEXTUALIZATION IN LARGE-SCALE SOCIAL NETWORKS

By

Rizwana Irfan

The Supervisory Committee certifies that this *disquisition* complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Samee U. Khan
Chair

Dr. Ying Huang

Dr. Scott C. Smith

Dr. Jacob Glower

Approved:

February 26, 2015
Date

Dr. Scott C. Smith
Department Chair

ABSTRACT

Social computing-based applications provide a coherent medium through which people can be interactive and socialize by developing a Web-based communication channel that integrates different Social Networking Services (SNSs) in the Social Networking Platforms (SNPs). Different SNSs, such as photo, audio, and video sharing, have emerged as an essential resources for the dissemination of information about the human interaction patterns. Most of the SNSs are integrated into a comprehensive and coherent paradigm called the Social Networking Platform (SNP). Most of the existing SNSs focused on content-based, media-based, and geo-location-based approach. The content-based SNSs allow the text-based interactions among individuals, such as communities, blogs, and social news. The media-based SNSs provide the social interaction through various multimedia formats, such as video and audio. Geo-location-based SNSs provide location-based social communication. However, all of the aforementioned techniques lack the semantic analysis which is the most integral and crucial part of the true understanding.

The goal of this dissertation is to incorporate the existing SNSs into the context-enriched information that provide the services customization based on the individual human characteristics, such as human preferences, and emotions. The computer interactive infrastructure can be enriched by leveraging information about the users' personal context (profile, preferences, attitude, and habits) that provides sophisticated context-aware services, such as semantic-based search and context-aware recommendations. The dissertation proposes *MobiContext*, a cloud-based Bi-Objective Recommendation Framework (BORF) for mobile social networks that generates real-time recommendation of venues for a group of mobile users.

The *MobiContext* utilizes multi-objective optimization techniques to generate personalized recommendations. To address the issues pertaining to cold start and data sparseness, the BORF performs data preprocessing by using the Hub-Average (HA) inference model. Moreover, the Weighted Sum Approach (WSA) is implemented for scalar optimization and an evolutionary algorithm (NSGA-II) is applied for vector optimization to provide optimal suggestions to the users about a venue. The dissertation also proposes a *SocialRec*, a context-aware recommendation framework that utilizes a rating sentiment inference approach to incorporate textual users' review into traditional collaborative filtering methods for personalized recommendations. The proposed framework utilizes semantic analysis scores on the users' contextual information to produce optimal recommendations.

ACKNOWLEDGMENTS

I am grateful to acknowledge and thank all those who assisted me in my graduate program at North Dakota State University. I would like to express my deepest appreciation and a bundle of thanks to my academic advisor Dr. Samee U. Khan. His guidance, support, and patience throughout my years as a graduate student are truly appreciated. Special thanks to my other graduate committee members, Dr. Ying Huang, Dr. Scott C. Smith, and Dr. Jacob Glower. I would also like to express my gratitude to COMSATS Institute of Information Technology, Pakistan for their financial support.

Last but not the least, I would like to thank all my colleagues at NDSU for their kind help and support during my PhD.

DEDICATION

I would like to dedicate this dissertation to my parents

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

1. INTRODUCTION

1.1. Overview

Social networking websites create new ways for engaging people belonging to different communities [1.20]. Social networks allow users to communicate with people exhibiting different moral and social values. The websites provide a very powerful medium for communication among individuals that leads to mutual learning and sharing of valuable knowledge [1.21]. The most popular social networking websites are Facebook, LinkedIn, and MySpace where people can communicate with each other by joining different communities and discussion groups. Social networking can solve coordination problems among people that may arise due to geographical distance [1.22], [1.23] and can increase the effectiveness of social campaigns [1.20], [1.23] by disseminating the required information anywhere and anytime. Social computing-based applications provide a coherent medium through which people can be interactive and socialize by developing a Web-based communication channel that integrates different Social Networking Services (SNSs) in the Social Networking Platforms (SNPs).

A Web-based social space termed as SNS is specifically designed for the end user-driven applications that enable communication, collaboration, and sharing of the knowledge though an assortment of a media [1.3]. Different SNSs, such as photo, audio, and video sharing, have emerged as an essential resources for the dissemination of information about the human interaction patterns. Roblyer's analysis about the current SNSs shows that the SNSs have supplanted TV as a popular medium for acquiring the information. According to Roblyer, 55% of Americans spend more time using SNSs than watching TV [1.2].

Most of the SNSs are integrated into a comprehensive and coherent paradigm called the Social Networking Platform (SNP). The SNPs provide a framework for different SNSs to

integrate and propagate. Moreover, SNPs allow users to establish social communications based on the mutual interests and cultural backgrounds [1.1]. The most popular SNPs are Facebook, LinkedIn, MySpace, Tumblr, Instagram, Google+, and Friendster. According to a recent statistical analysis, Facebook has more than 1 billion active subscribers worldwide and Google+ has more than 250 million active users [1.4], [1.2].

In social computing, interactivity patterns of the human behaviors are based on the contents. Most of the work presented in the past focused on: (a) content-based, (b) media-based, and (c) geo-location-based, (d) context-aware SNSs [1.5], [1.6], [1.1], [1.7], [1.8]. The content-based SNSs allow the text-based interactions among individuals, such as communities, blogs, and social news. The media-based SNSs provide the social interaction through various multimedia formats, such as video and audio. Geo-location-based SNSs provide location-based social communication. The context-aware SNSs improve the quality of interaction by providing service oriented architecture for the social computing. We believe that the media, geo-location, and context based SNSs can improve the traditional text-based interaction of the content-based SNSs as presented in Figure 1.

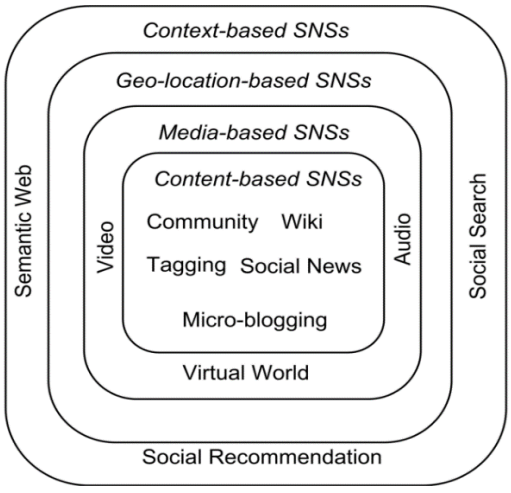

Figure 1. Social networking services framework

In conventional social interactive environments, such as Facebook, LinkedIn, and MySpace, computers are not capable of acquiring the information based on common intelligence [1.16]. An integration of the contextual information with interactive computing can be a promising solution for the development of effective intelligent social communicational services [1.17], [1.18]. Following the formal definition of contextual information formulated by Anind et al. in [1.17], the term "context" refers to the relevant information that can be used for the categorization of various attributes and situations of entities, where the entities can be a place, person, or object.

There are two main categories of contextual information, namely: (a) physical contextual information and (b) logical contextual information, as presented in Figure 2. Physical contextual information can be acquired by using hardware sensors, whereas the logical contextual information can be obtained by analysing the human habits, attitudes, and preferences [1.19].

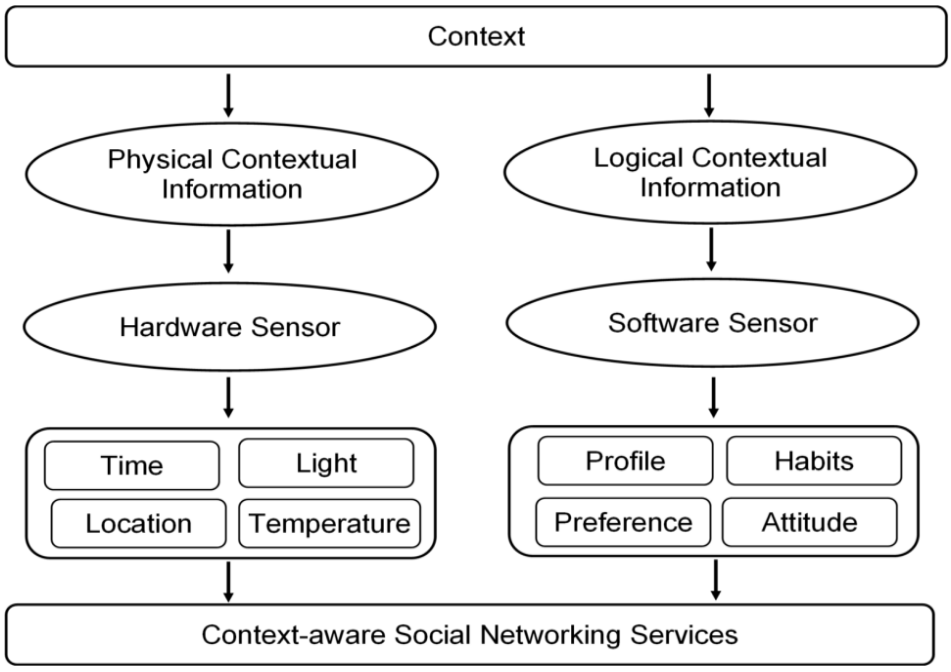

Figure 2. Context-aware social networking services

As a result of the ease of application and large availability of hardware sensors, the majority of context-aware systems use physical contextual information. However, the successful implementation of logical contextual information in social computing is still a challenging task for researchers due to the high complexity of human behaviours [1.18], [1.17].

1.2. Motivation

All of the techniques discussed in introduction Section lack the semantic analysis which is the most integral and crucial part of the true understanding. In [1.4], [1.1], [1.9], [1.2], [1.10], the authors defined and compared different SNSs. Ref. [1.9], [1.4], [1.11] presented different statistical analyses of the popular SNSs, and in [1.2], the author discussed different types of the content-based SNSs. However, all of the abovementioned studies overlooked the importance of context-aware computing in the SNSs. Gartner, the most prestigious information technology and advisory company of the USA, predicted that the next generation of computers will be the context-aware computers [1.12]. Mark Weiser, chief scientist at Xerox PARC, emphasized the integration of context-enriched services, such as location and social attributes to anticipate the end users' requirements. The context-aware computing provides the services customization based on the individual human characteristics, such as human preferences, mood, behaviours, and emotions [1.12]. Moreover, the context-aware computing improves the quality of interaction by providing service oriented architecture for the social computing [1.13].The importance of context-based SNSs has increased in the past few decades, yet no substantial amount of research has been performed in this burgeoning area of the social computing [1.14], [1.13], [1.15]. The computer interactive infrastructure can be enriched by leveraging information about the users' personal context (profile, preferences, attitude, and habits) that provides sophisticated context-aware services, such as semantic-based search and recommendations. Recommendation systems

are increasingly emerging as an integral component of e-business applications [1.1]. For instance, the integrated recommendation system of Amazon provides customers with personalized recommendations for various items of interest. Recommendation systems utilize various knowledge discovery techniques on a user's historical data and current context to recommend products and services that best match the user's preferences. Recently, CARSs have been discussed extensively in scientific research literature, particularly the location recommendation services for various mobile devices [40], shopping assistants [1.12], and conference assistants [1.12], [1.14].

In Content-based SNS, people generally use unstructured or semi-structured language for communication. In everyday life conversation, people do not care about the spellings and accurate grammatical construction of a sentence that may leads to different types of ambiguities, such as lexical, syntactic, and semantic [1.21]. Therefore, extracting logical patterns with accurate information from such unstructured form is a critical task to perform. Text mining can be a solution of above mentioned problems. Due to the increasing number of readily available electronic information (digital libraries, electronic mail, and blogs), text mining is gaining more importance. Text mining is a knowledge discovery technique that provides computational intelligence [1.21]. The technique comprises of multidisciplinary fields, such as information retrieval, text analysis, natural language processing, and information classification based on logical and non-trivial patterns from large data sets. In [1.24], the authors defined text mining as an extension of data mining technique. The data mining techniques are mainly used for the extraction of logical patterns from structured database. Text mining techniques become more complex as compared to data mining due to unstructured and fuzzy nature of natural language text [1.24].

1.3. Contributions

The objective of our research is to architect content-based and context-based social network services that provide computer-mediated communication that promotes the interaction among individuals.

1.3.1. Context-aware Recommendation Systems as Social Networking Services

In recent years, recommendation systems have seen significant evolution in the field of knowledge engineering. Most of the existing recommendation systems based their models on collaborative filtering approaches that make them simple to implement. However, performance of most of the existing collaborative filtering-based recommendation system suffers due to the challenges, such as: (a) cold start, (b) data sparseness, and (c) scalability. Moreover, recommendation problem is often characterized by the presence of many conflicting objectives or decision variables, such as users' preferences and venue closeness. In this paper, we proposed MobiContext, a cloud-based Bi-Objective Recommendation Framework (BORF) for mobile social networks. The MobiContext utilizes multi-objective optimization techniques to generate personalized recommendations. To address the issues pertaining to cold start and data sparseness, the BORF performs data preprocessing by using the Hub-Average (HA) inference model. Moreover, the Weighted Sum Approach (WSA) is implemented for scalar optimization and an evolutionary algorithm (NSGA-II) is applied for vector optimization to provide optimal suggestions to the users about a venue.

1.3.2. The Content-based SNSs

The content-based SNSs allow the text-based interactions among individuals, such as communities, blogs, and social news. Social networking websites, such as Facebook are rich in texts that enable user to create various text contents in the form of comments, wall posts, social

media, and blogs. Due to ubiquitous use of social networks in recent years, an enormous amount of data is available via the Web. Application of text mining techniques on social networking websites can reveal significant results related to person-to-person interaction behaviours. Moreover, text mining techniques in conjunction with social networks can be used for finding general opinion about any specific subject, human thinking patterns, and group identification in large-scale systems [1.33]. For the past few years there has been a lot of research in the area of text mining. In the scientific literature [1.20], [1.38], [1.37], various text mining techniques are suggested to discover textual patterns from online sources. In [1.34], the authors restrict the analysis to techniques that are specifically associated with text document classification. Brucher stated various clustering based approaches for document retrieval and compared different clustering techniques for logical pattern extraction from unstructured text, but most of the techniques presented in the papers are not recent [1.34]. In [1.35], the authors proposed a new model for textual categorization to capture the relations between words by using WordNet ontology [1.36]. The proposed approach maps the words comprise of same concepts into one dimension and present better efficiency for text classification. In [1.36], the authors indicated a best practice in information extraction process based on semantic reasoning capabilities and highlighted various advantages in terms of intelligent information extraction. The author explained the suggested methods, such as query expansion and extraction for semantic based document retrieval, but did not mention any results associated with the experiments. In [1.37], the author introduced general text mining framework to extract relevant abstract from large text data of research papers. However, the proposed approach neglected the semantic relations between words in sentences.

Most of the scientific literature [1.36], [1.37], [1.36] focuses on specific techniques of text mining for information extraction from text documents. However, a thorough discussion is lacking on the actual analysis of different text mining approaches. Most of the surveys emphasize on the application of different text mining techniques on unstructured data but do not specifically target the datasets in social networking websites. Moreover, the existing research papers cover the text mining techniques without mentioning the pre-processing phase [1.35], 1.36] that is an important phase for the simplification of text mining process. In contrast, this survey attempts to address all the above mentioned deficiencies by providing a focused study on the application of all (classification and clustering) text mining techniques in social networks where data is unstructured. Ontologies overcome the difficulties raised by monolithic, isolated knowledge systems by specifying a content-specific agreement to facilitate the knowledge sharing and reuse in a specific domain. Moreover, ontologies comprise: (i) objects, (ii) concepts, and (iii) hierarchical relationships between concepts and objects to acquire semantics in a specific field, such as medicine, academia, and engineering [1.31], [1.30], [1.32]. For instance, medical ontology contains the basic concepts related to treatments of various diseases and clinical procedures that facilitate the propagation of standard medical terminology in the healthcare systems.

Our survey focuses on three major aspects: (i) ontology learning techniques, (ii) the ontology learning process, and (iii) the implication of various ontology learning techniques in the ontology learning process. Moreover, the discussion presents the ontology-based text mining architecture as an example of ontology learning implication into the field of text mining, the most promising research area for logical interpretation of the text corpora. Furthermore, we review the major issues and challenges in the ontology learning process. Moreover, this research

provides several possible future research direction especially in the field of intelligent text analysis in large-scale systems, such as social Web.

1.4. List of Publications

Some of the contributions presented in this dissertation have appeared in the following publications:

1. R. Irfan, G. Bickler, S. U. Khan, J. Kolodziej, H. Li, D. Chen, L. Wang, K. Hayat, S. A. Madani, B. Nazir, I. A. Khan, and R. Ranjan, "Survey on Social Networking Services," IET Networks. (Forthcoming.)
2. R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, N. Tziritas, C.-Z. Xu, A. Y. Zomaya, A. S. Alzahrani, and H. Li, "A Survey on Text Mining in Social Networks,"Knowledge Engineering Review. (Forthcoming.)
3. R.Irfan, Samee U. Khan, Camelia Chira, Pavan Balaji, Fan Zhang, Rajiv Ranjan, Sajjad Madani, Dan Chen " Survey of Ontology Learningn in text Mining" IET Networkd (Submitted)
4. K. Bilal, S. U. R. Malik, O. Khalid, A. Hameed, E. Alvarez, V. Wijaysekara, R. Irfan, S. Shrestha, D. Dwivedy, M. Ali, U. S. Khan, A. Abbas, N. Jalil, and S. U. Khan, "A Taxonomy and Survey on Green Data Center Networks,"Future Generation Computer Systems. (Forthcoming.)
5. Tahir Maqsood, R. Irfan, Sajjad A. Madani, Samee U. Khan, "Scalalability Issues in Social Networks", (Submitted)
6. F. A. Shah, R. Irfan, S. U. Khan, S. A. Madani, T. Iqbal, : A Survey on Affective Tutoring systems", (Submitted)

7. R. Irfan, Osman Khalid, Usman Shahid, “Context-aware Recommendation Systems Based on Bi-Objective Optimization Techniques, (Submitted).
8. R. Irfan and S. U. Khan, "Scalable Services in Social Network Services," IEEE Technical Committee on Scalable Computing Blog, September 03, 2012.

1.5. Dissertation Outline

The dissertation is organized as follows. In Chapter 2, we present the background and related literature. Chapter 3 presents an empirical study on ontology learning and its implication in text mining field. In Chapter 4 we present a venue recommendation system for mobile social networks. Chapter 5 presents a context-aware recommendation framework that utilizes a rating inference approach to incorporate textual users’ review into traditional collaborative filtering methods for personalized recommendations. Chapter 6 presents and conclusions with future research directions.

1.6. References

[1.1] Irwin, K.: ‘Introduction to Social Computing’. Proc. Int. Conf. Database System for Advanced Applications, 2010, pp. 482–484

[1.2] Roblyer, M.D., Michelle, M., Webb, M., Herman, J., and Witty, J.V.: ‘Findings on Facebook in Higher Education: A Comparison of Collage Faculty and Students Uses and Perception of Social Networking Sites’, The Internet and Higher Education, June 2010, 1, (3), pp. 134-140

[1.3] Nov, O., and Ye, C.: ‘Community Photo Sharing: Motivational and Structural Antecedents’. Proc. Int. Conf. Information Systems (ICIS), 2008, pp. 1-10S. U. Khan, “Mosaic-Net: A Game Theoretical Method for Selection and Allocation of Replicas in Ad Hoc Networks,” Journal of Supercomputing, vol. 55, no. 3, pp. 321-366, 2011.

[1.4] Ahn, Y.Y., Han, S., Kwak, H., Eom, Y.H., Moon, S., and Jeong, H.: 'Analysis of Topological Characteristics of Huge Online Social Networking Services'. Proc. Int. Conf. World Wide Web, 2007, pp. 835-844

[1.5] Zheng, Y., Li, Q., Chen, Y., Xie, X., and Ma, W.Y.: 'Understanding Mobility Based on GPS Data'. Proc. Int. Conf. Ubiquitous Computing, 2008, pp. 312-321

[1.6] Zheng, Y., and Xie, X.: 'Learning Location Correlation User-Generated GPS Trajectories'. Proc. Int. Conf. Mobile Data Management (MDM), 2010, pp. 27-32

[1.7] Zheng, Y., Zhang, L., Ma, Z., Xie, X., Ma, W.Y.: 'Recommending Friends and Location based on Individual Location History', ACM Transactions on the Web (TWEB), 2011, 5, (1), pp. 12-45

[1.8] Zheng, Y., Chen, Y., Xie, X., and Ma, W.Y.: 'Understanding Transportation Modes Based on GPS Data for Web Application', ACM Transaction on the Web, Jan 2010, 4, (1), pp. 1-36

[1.9] Sorensen, L.: 'User Managed Trust in Social Networking Comparing Facebook, MySpace and LinkdIn'. Proc. Inter. Conf. Wireless Communication, Vehicular Technology, Information Theory and Aerospace and Electronic System Technology, (Wireless VITAE '09), 2009, pp. 427-431

[1.10] Koren, Y.: 'Factor in the Neighbors: Scalable and Accurate Collaborative Filtering', ACM Transaction on Knowledge Discovery from Data (TKDD), 2010, 4, (1), pp. 34-45P.

[1.11] Chiu, C.M., Hsu, M.H., and Wang, E.G.: 'Understanding Knowledge Sharing in Virtual Communities: An Integration of Social Capital and Social Cognitive Theories', Decision Support System, Dec 2006, 42, (3), pp. 1872-1888

[1.12] http://www.gartner.com/technology/research/top-10-technology-trends/ Gartner, "Top Ten Strategic Technology Trends for 2012", Special Report, Released on Jan 2012, accessed June 13, 2012S.

[1.13] http://blog.nielsen.com/nielsenwire/wp-content/uploads/2009/03/overview-of-home-internet-access-in-the-us-jan-6.pdf. Nielsen, An Overview of Home Internet Access in the US, Technical report, accessed June 13, 2012

[1.14] Chi, E.H.: 'Information Seeking can be Social', Computer, March 2009, 42, (3), pp. 42-46

[1.15] Baldauf, M., Dustdar, S., and Rosenberg, F.: 'A Survey on Context-Aware Systems', International Journal of Ad Hoc and Ubiquitous Computing, June 2007, 2, (4), pp. 263-277

[1.16] Anderson, J., Diaz, C., Bonnean, j., Stajano. F.: 'Privacy-enabling social Netwrking over Untrusted Networks', Proc. Int. Conf. Online Social Networks, 2009, pp. 1-6

[1.17] Bettini, C., Brdiczka, O., Henricksen, K., Indulska, J., Nicklas, D., and Ranganathan, A.: 'A Survey of Context Modeling and Reasoning Techniques', Pervasive and Mobile Computing, 2010, 6, (2), pp. 161-180

[1.18] Kulkarni, D., and Tripathi, A.: 'A Framework for Programming Robust Context-Aware Applications', IEEE Transaction on Software Engineering, April 2010, 36, (2), pp. 184-197

[1.19] Castro, A.G., Labarga, A., Garcia, L., Giraldo, O., Montana, C., and Bateman, J.A.: 'Semantic Web and Social Web Heading towards Living Documents in the Life Science', Web Semantics: Science, Services and Agent on the World Wide Web, Jul 2010, 8, (2), pp. 155-162

[1.20] Baumer, E. P. S., Sinclair, J. & Tomlinson, B. 2010. America is like metamucil: Fostering critical and creative thinking about metaphor in political blogs. In Proceedings of 28th International Conference on Human Factor in Computing Systems (CHI 2010) ACM, Atlanta, GA, USA, 34-45.

[1.21] Sorensen, L. 2009. User managed trust in social networking comparing facebook, myspace and linkdin. In Proceed- ings of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic System Technology, (Wireless VITAE 09), Denmark, 427-431.

[1.22] Evans, B. M., Kairam, S. & Pirolli, P. 2010. Do your friends make you smarter: An analysis of social strategies in online information seeking. Information Processing and Management, 46 (6), 679-692.

[1.23] Li, J., Li, Q., Liu, C., Khan, S. U. & Ghani, N. 2012. Community-based collaborative information system for emergency management. Computers and Operations Research. (To appear.)

[1.24] Liu, F. & Lu, X. 2011. Survey on text clustering algorithm. In Proceedings of 2nd International IEEE Conference on Software Engineering and Services Science (ICSESS), China, 901-904.

[1.25] Kano,Y., Baumgartner, W. A., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L. & Tsujii, T. 2009. Data Mining: Concept and Techniques. Oxford Journal of Bioinformatics, 25(15), 1997-1998.

[1.26] Hsieh, S., Lin, H., Chi, N., Chou, K., and Lin, K.: 'Enabling the Development of Base Domain Ontology through Extraction of Knowledge from Engineering Domain Handbooks', Advanced Engineering Informatics, 2011, 2(25), pp. 288-296

[1.27] Marinica, C., Guillet, F.: ‘Knowledge-Based Interactive Post-mining of Association Rules using Ontologies’, Knowledge and Data Engineering, 2010, 6(22), pp.784-797

[1.28] Khalida, B., Adil. T.: ‘Lightweight Domain Ontology Learning from Texts: Graph Theory-based Approach using Wikipedia’, International Journal of Metadata, Semantics and Ontologies, 2014, 9(2), pp. 83-90

[1.29] Wong, W., Liu, W., and Bennamoun, M.: ‘Ontology Learning from Text: A Look Back and into the Future’, ACM Journal Computing Survey, 2012, 44(4), pp. 12-32

[1.30] Marinica, C., Guillet, F.: ‘Knowledge-Based Interactive Post-mining of Association Rules using Ontologies’, Knowledge and Data Engineering, 2010, 6(22), pp.784-797

[1.31] Sergeja, V., and Zoran, B.: ’Ontology-based Multi-Label Classification of Economic Article’, Computer Science and Information Systems (ComSIS), 2011, 1(1), pp. 101-119

[1.32] Turney, P., and Pantel, P.: ‘From Frequency to Meaning: Vector Space Models of Semantics’, Journal of Artificial Intelligence, 2014, 37, pp. 141-188

[1.33] Aggarwal, C. 2011. Text mining in social networks. In Social Network Data Analytics. 2nd edn. Springer, 353-374.

[1.34] Baumer, E. P. S., Sinclair, J. & Tomlinson, B. 2010. America is like metamucil: Fostering critical and creative thinking about metaphor in political blogs. In Proceedings of 28th International Conference on Human Factor in Computing Systems (CHI 2010) ACM, Atlanta, GA, USA, 34-45.

[1.35] Durga, A. K. & Govardhan, A. 2011. Ontology based text categorization-telugu document. International Journal of Scientific and Engineering Research, 2(9), 1-4.

[1.36] Xu, X., Zhang, F. & Niu, Z. 2008. An ontology-based query system for digital libraries. In Proceedings of IEEE, Pacific-Asia Workshop on Computational Intelligence and Industrial Application, Wuhan, 222-226.

[1.37] Tekiner, F., Aanaiadou, S.,Tsuruoka, Y. & Tsuji, J. 2009. Highly scalable text mining parallel tagging application. In Proceedings of IEEE 5th International Conference on Soft Computing, Computing with Words and Perception in System Analysis, Decision and Control (ICSCCW), China, 1-4.

[1.38] Ringel, M. M., Teevan, J. & Panovich, K. 2010. What do people ask their social networks, and why: A survey study of status message question & answar behavior. In Proceedings of International Conference on Human Factors in Computing Systems (CHI 10), Atlanta, GA, USA, 56-62.

2. BACKGROUND AND RELATED WORK

This chapter presents the background as well as the literature survey on recent works related to the topics investigated throughout this dissertation.

2.1. Background

The social computing, such as Social Networking services (SNSs) and social Networking Platforms (SNPs) provide a coherent medium through which people can be interactive and socialize. The SNP is a Web-based social space, specifically designed for end user-driven applications that facilitate communication, collaboration, and sharing of the knowledge through a variety of SNSs, such as text, video, and audio streams. In the conventional SNPs, such as Facebook, LinkedIn, and MySpace, computers are not capable of acquiring the information based on the common intelligence and human behavior. This chapter provides a comprehensive overview of the current SNSs and discusses different possibilities of incorporating the existing SNSs into the context-aware techniques that include ontologies, text mining, and social recommendations. The context-aware computing provides services customization based on the individual human characteristics, such as human preferences, mood, behaviors, and emotions. The Integration of contextual information with SNSs can be more useful and productive for the development of the intelligent social communicational services.

2.2. Context-based Social Networking Services

Following the formal definition of contextual information formulated by Anind et al. in [2.1], the term “context” refers to the relevant information that can be used for the categorization of various attributes and situations of entities, where the entities can be a place, person, or object. The context-aware SNSs provide an appropriate platform for the integration of physical and

logical contextual information that can be gleaned from tagging a picture or joining different communities.

2.3. Context-aware Recommendation Systems as Social Networking Services

The Web-based recommendation systems have evolved at a prodigious rate over the past few decades. A recommendation algorithm garners user interests as an input and creates a list of recommendations. Moreover, the recommendation algorithms help online users to avoid information overload by filtering the information [2.2], [2.3]. For instance, Amazon.com can be considered a good example of such systems, where recommendations of the products are based on a customer's interests. The basic recommendation systems are classified into two main categories, namely: (a) content-driven recommendation systems and (b) collaborative-filtering-based recommendation systems [2.4]. The content-driven recommendation systems recommend products based on the product description and customer's interests. The implementation of content-based recommendation systems is based on a keyword approach. In the keyword approach, the importance of any word in the document can be evaluated by using different weighted measure techniques, such as :(a) Term Frequency/Inverse Document Frequency (TF-IDF), (b) Bayesian classifiers, (c) clustering, and (d) Decision Trees (DT) [2.5], [2.6], [2.7]].

A Collaborative Filtering (CF) approach predicts recommendations by evaluating an item's purchase history and an item's grading criteria through similar users [2.8]. The CF structure presents the problem as a two-dimensional matrix comprised of pairwise values of the users and items [2.9]]. The CF approach has attracted much interest due to the ability of exploring the complex data patterns without extensive data collection [2.9]. Several successful commercial systems, such as Amazon, Netflix, and LastFM [38] use CF-based recommendation systems to generate the recommendations about items, such as news, books, and movies.

However, according to Fengkun in [2.8], CF is unable to identify the neighbors as friends or strangers. While making decision, people usually rely on recommendations from a friend rather than a stranger [2.8]. The substantial amount of work has been done in the field of recommendation systems. However, most of the existing approaches focus on recommending items to users and users to items and do not consider contextual information, such as place and time. Therefore, context-aware recommendation systems have been developed for the integration of the contextual information into the recommendation systems as explained in the subsequent text.

Various SNPs, such as MySpace, Facebook, and LinkedIn pioneered the combination of the social relationship information of users with the CF to generate advanced CF-based recommendation systems [2.10]. Similarly, Liu and Lee in [2.8] proposed a "CF with friends" approach that selects nearest neighbors based on the social network information of the user for the product recommendations. Moreover, Ref. [2.8] used the social network data for the recommendations using a three-fold process: (1) collect data about the preferences and social relationships of the users, (2) develop strategies to select the nearest neighbor based on the social relationships, and (3) generate recommendations. Liu and Lee presented an innovative concept to enhance the recommender system. However, the analysis presented in their paper is limited to a single Cyworld Website in South Korea and most of the experiments were conducted on the social Web service simulator, which may not reflects the realistic social Web-based scenario for the users located in diverse regions [39]. Another successful application of combining CF with social network information has been observed in [2.10, [2.11].

Another context-aware recommendation systems based approach has been suggested by Sieg et al. that present a hybrid semantic Web system that integrates the CF with ontologies

[2.12], [2.13]]. The authors introduced an ontological-profile of users based on their behavior. The user profiles are updated based on the interactions and relationships among the users, defined by the ontologies. In the recommendation process, the ontology-based user profiles are compared to generate the semantic neighborhoods. However, extraction of the contextual information using ontologies is a complex process in the real life scenarios due to the multifaceted nature of human behavior.

The concept of context-aware recommendation systems seems to be a promising solution for future generations of Social Semantic Networking Services (SSNSs) and can be easily extended by developing new modules and tools [2.14, 2.15]. In scientific literature context-aware venue recommendation systems have been investigated in mobile social networks as presented in the Figure 3.

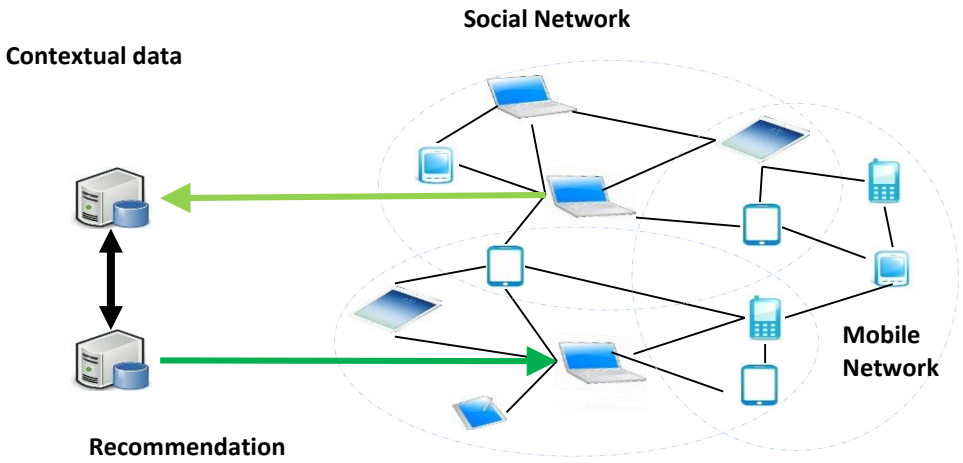

Figure 3. Social networking services evolution

The existing context-aware venue recommendation systems approaches can be categorized as [2.20]: (a) trajectory based, (b) explicit rating based, and (c) check-in based approaches. Trajectory based approaches utilize information about a user's visit sequence to

various locations, the paths selected, and the duration of stays. Doytsher et al. [2.16] proposed a trajectory-based graphical model that keeps track of frequently traveled routes by users and recommend best route to a new user. The authors in [2.17] mine GPS trajectories data to extract most popular locations based on users' travel sequences. Although the aforementioned approaches suggests locations based on users' past trajectories, they are unable to distinguish the places in terms of their categories, which we performed in our proposed MobiContext framework.

Many online social services, such as Yelp (yelp.com) and Yellow pages (yellowpages.com) allow users to rate the visited locations. Rating-based venue recommendation systems utilize the existing ratings' data to recommend people with most popular venues or travel routes in a city. The authors in [2.18] proposed models based on collaborative filtering that take into account users' existing ratings to generate personalized venue recommendations. The aforementioned approaches may closely capture users' preferences, but are not scalable enough to simultaneously process huge volumes of real-time data. Moreover, they also suffer from data sparseness issues due to limited number of entries within the user-rating matrix.

Most of the above mentioned approaches have designs built on (memory based) CF models, which enables these approaches to depict a user's future preferences based on his/her past entries. However, these approaches suffer from scalability issues due to large number of similarity computations on user-venue matrix during online recommendation process. Moreover, such approaches also suffer from data sparseness and cold start problems, as there are very few users who have visited large number of venues. Furthermore, these approaches do not provide a solution to the multi-objective. To address these limitations, our proposed cloud based recommendation framework, MobiContext presents a solution for scalability and data sparseness.

2.3.1. Ontologies

Ontologies establish a common understanding between humans and machines of a situation, event, or object. Valls et al. defined ontologies as meta-information that provides information about the inter-document relations in a machine executable format [2.19].

The syntax and semantic based interpretation of a textual document in the Web-based systems is crucial for the retrieval of up-to-date, consistent, and accurate information. Web-based systems, such as Webmail, e-commerce, and wikis require different ontologies for the correct interpretation of logical relationships between semi-structured texts [2.20]. However, building a universal ontology is still difficult, especially with a large number of users exhibiting diverse backgrounds.

2.4. Text Mining in Content-based SNSs Using Classification

Content-based SNSs are primarily textual and provide basic computer-mediated communication that promotes the text-based interaction among individuals. However, the functionality of the content-based SNSs can be further extended by incorporating media, geo-location, and context-based SNSs. The contents of content-based SNSs are generated by the users. Therefore, the successful implementation of content-based SNSs depends on the willingness of the contributors to share information [2.45]. Basic content-based services are blogs, Wikipedia, micro-blogging, social news, tagging, and chatting.

In content-based SNSs, people generally use unstructured or semi-structured language for communication. In everyday life conversation, people do not care about the spellings and accurate grammatical construction of a sentence that may leads to different types of ambiguities, such as lexical, syntactic, and semantic [2.21]. Therefore, extracting logical patterns with accurate information from such unstructured form is a critical task to perform.

Text mining can be a solution of above mentioned problems. Due to the increasing number of readily available electronic information (digital libraries, electronic mail, and blogs), text mining is gaining more importance. Text mining is a knowledge discovery technique that provides computational intelligence [2.21]. The technique comprises of multidisciplinary fields, such as information retrieval, text analysis, natural language processing, and information classification based on logical and non-trivial patterns from large data sets. In [2.22] the authors defined text mining as an extension of data mining technique. The data mining techniques are mainly used for the extraction of logical patterns from structured database. Text mining techniques become more complex as compared to data mining due to unstructured and fuzzy nature of natural language text [2.23].

Supervised learning or classification is the process of learning a set of rules from a set of examples in a training set. Text classification is a mining method that classifies each text to a certain category [2.24]. Classification can be further divided into two categories: (a) machine learning based text classification and (b) ontology based text classification [2.25] and is illustrated in Figure 4.

2.4.1. Machine Learning-based Text Classification

Machine Learning based Text Classification (MLTC) comprises of quantitative approaches to automate Natural Language Processing (NLP) that uses machine learning algorithms. Preferred supervised learning techniques for text classification are described in the subsequent text.

2.4.1.1. Rocchio Algorithm

Different words with similar meanings in a natural language are termed as Synonymy. Synonymy can be addressed by refining the query or document using the relevance feedback

method. In the relevance feedback method, the user provides feedback that indicates relevant material regarding the specific domain area. The user asks a simple query and the system generates initial results in response to the query. The user marks the retrieved results as either relevant or irrelevant. Based on the users marked results the algorithm may perform better. The relevance feedback method is an iterative process and plays a vital role by providing relevant material that tracks user information needs [2.22].

2.4.1.2. Instance-based Learning Algorithm

Instance based learning algorithms (also known as lazy algorithms) are based on the comparison between new problem instances and instances already stored during training [2.26]. On arrival of a new instance, sets of related instances are retrieved from the memory and further processed so the new instance can be classified accordingly. Algorithms exhibiting instance based learning approaches are described in the subsequent text.

K-Nearest Neighbour (K-NN) algorithm is a form of instant based learning. The algorithm categorizes similar objects based on the closest feature space in the training set. The closest feature space may be determined by measuring the angle between the two feature vectors or by calculating the Euclidean distance between the vectors. For more details, we encourage the readers to browse [2.26].

Case Based Reasoning (CBR) comprises of three basic steps: (1) classification of a new case by retrieving appropriate cases from data sets, (2) modification of the extracted case, and (3) transformation of an existing case [2.27]. Textual Case Based Reasoning (TCBR) primarily deals with textual knowledge sources in making decisions. A novel textual case-based reasoning system, named SOPHIA-TCBR has been detailed in [2.28] for organizing semantically related textual data into a group. Ref. [2.28] stated better results of knowledge discovery in the

SOPHIA-TCBR system. However, in the TCBR approach, extracting similar cases and representing knowledge without losing key concepts with low knowledge engineering overhead are still challenging issues for researchers [2.28].

2.4.1.3. Decision Trees and Support Vector Machine

Relationships, attributes, and classes in ontology can be structured hierarchically as taxonomies [2.27]. The process of constructing lexical ontology by analysing unstructured text is termed as ontology refinement. Decision Tree (DT) is a method to semantically describe the concepts and the similarities between the concepts [2.27]. Different algorithms of decision tree are used for classification in many application areas, such as financial analysis, astronomy, molecular biology, and text mining. As text classification depends on a large number of relevant features, an insufficient number of relevant features in a decision tree may lead to poor performance in text classification [2.27].

Support Vector Machine (SVM) algorithm is used to analyse data in classification analysis. In contrast to other classification methods, SVM algorithm uses both negative and positive training datasets to construct a hyper plane that separates the positive and negative data. The document that is closest to decision surface is called support vector [2.29]. For detailed description refer to [2.29].

2.4.1.4. Artificial Neural Networks

Artificial Neural Networks (ANN) are parallel distributed processing systems specifically inspired by the biological neural systems [2.30].The network comprises of a large number of highly interconnected processing elements (neurons) working together to solve any specific problem. Due to their tremendous ability to extract meaningful information from a huge set of data, neurons have been configured for specific application areas, such as pattern recognition,

feature extraction, and noise reduction. In the neural network, connection between two neurons determines the influence of one neuron on another, while the weight on the connection determines the strength of the influence between the two neurons [2.30].

There are two basic categories of learning methods used in neural networks: (a) supervised learning and (b) unsupervised learning. In supervised learning, the ANN gets trained with the help of a set of inputs and required output patterns provided by an external expert or an intelligent system. Different types of supervised learning ANNs include: (a) back propagation and (b) modified back propagation neural networks [2.32]. Major application areas of supervised learning are pattern recognition and text classification [2.30] [2.31] . In unsupervised learning (clustering), the neural network tends to perform clustering by adjusting the weights based on similar inputs and distributing the task among interconnected processing elements [2.32].

The field of text mining is gaining popularity among researchers because of enormous amount of text available via Web in the form of blogs, comments, communities, digital libraries, and chat rooms. ANN can be used for the logical management of text available on Web. Jo proposed a new neural network architecture for text categorization with document presentation called Neural Text Categorizer (NTC) [2.30]. NTC comprises of three layers: (a) input layer, (b) output layer, and (c) learning layer. Input layer is directly connected with output layer, whereas learning layers determine the weights between input and output layer. The proposed approach can also be use for organizing the text in social networks [2.30].

2.4.1.5. Genetic Algorithms

A Genetic Algorithm (GA) is a heuristic search that simulates the natural environment of biological and genetic evolution [2.32] [2.33]. Multiple solutions of a problem are presented in the form of a genome. The algorithm creates multiple solutions and applies genetic operators to

determine the best offspring. Genetic algorithms are widely used to solve optimization problems. Therefore, researchers are trying to use the utility of genetic algorithms in social networking websites [2.32], [2.34].

A genetic algorithm was used for feature selection and termed weight method in [2.35] for assigning weights to each concept in the document on the basis of relevant topics. Weighted Topic Standard Deviation (WTSD) was the proposed formula used to present the concentration of a topic in a document as a fitness function. As the process is recursive, an end function needs to be specified based on monitoring the improvement of results in the consecutive generations. In [2.35], the authors revealed better results by using a genetic algorithm for text classification.

2.4.2. Ontology-based Text Classification

Statistical techniques for document representation (as described in Section 3.1) are not sufficient because the statistical approach neglects the semantic relations between words [2.32]. Consequently, the learning algorithm cannot identify the conceptual patterns in the text [2.32]. Ontology can be the solution of the problems by introducing explicit specification of conceptualization based on concepts, descriptions, and the semantic relationships between the concepts [2.35] [2.36]. Ontology represents semantics of information and is categorized as: (a) Domain Ontology (DO) consists of concepts and relationship of the concepts about a particular domain area, such as biological ontology or industrial ontology and (b) Ontology Instance (OI) related with automatic generation of web pages [2.32].

Basic components of ontology include: (a) classes, (b) attributes, (c) relations, (d) function terms, and (e) rules [2.37]. Ontology needs to be specified formally [2.32]. Formal relation can be represented as: (a) classes and (b) instances [2.35]. Ontology based languages are declarative languages and generally express the logic of computation based on either first-order

logic or description logic. For instance, the W3C organization introduced standardized Ontology Web Language (OWL) that supports interpretability of language by providing additional vocabulary with formal semantics [2.25]. Common Logic (CL) [2.37] and Semantic Application Design Language (SADL) [2.37] are the popular ontology based languages commonly used for semantic evaluation of data sets available in social networking websites.

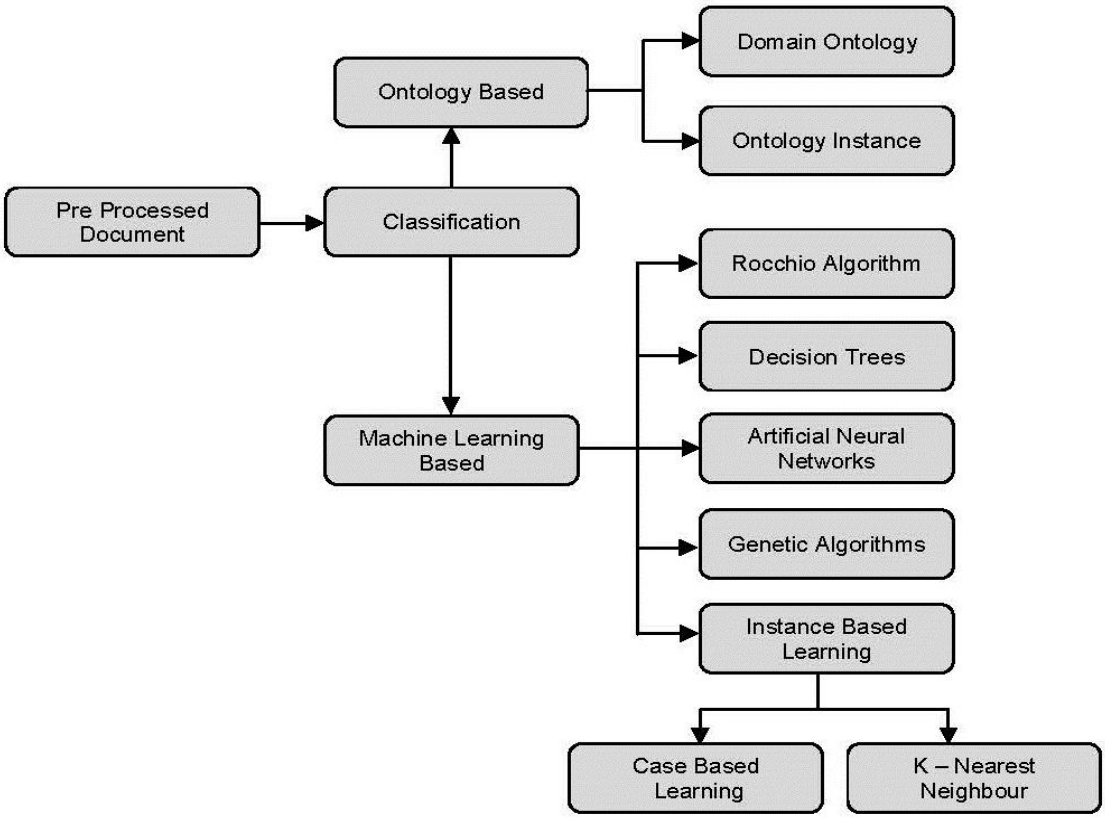

Figure 4. Text mining using classification

On-line information usually resides in digital libraries in the form of on-line books, conference, and journal papers. In digital libraries, searching techniques are based on a traditional keyword matching approach that may not satisfy requirements of users due to lack of semantic reasoning capabilities. Xu recommended an ontology-based digital library system that

analyzed the query with respect to semantic meanings and revealed better results when compared with traditional keyword based searching approach [2.25]. However, semantic analysis is computationally expensive and challenging for researchers especially for large text corpora, such as text data in social networking websites [2.25].

2.4.3. Hybrid Approach

Different classification algorithms have been used for text classification and analysis. However, literature [2.38], [2.39], [2.40], [2.41] shows that the combination of different classification algorithms (hybrid approach) provides better results and increased text categorization performance instead of applying a single pure method. The result of applying hybrid approach to large text corpora heavily depends on the test data sets. Therefore, there is no guarantee that a high level of accuracy acquired by one test set will also be obtained in another test set. Moreover, for better performance of the hybrid approach, several parameters need to be defined or initialized in advance. Table 1 provides an overview of different hybrid approaches used for text classification that can be further used for the text analysis in social networking. However, selecting the classification approach for text analysis in social networks totally depends on the dataset and nature of the problem being investigated [2.38].

The result of the analysis shows that SVM and ANN performed well in several comparisons. The main purpose of the comparison of hybrid approach is to highlight the applicability of different classification algorithms and complement their limitations [2.39].

Table 1. Comparison of hybrid approaches

Authors	Hybrid Approaches						Success Rate
	ANN	RA	DT	SVM	K-NN	GA	
(Miao at al. 2009)	No	Yes	No	No	Yes	No	83.8%
(Wu 2009)	Yes	No	Yes	No	No	No	93.67%
(Aci et al. 2010)	No	No	No	No	Yes	Yes	75.52%%
(Gazzah & Ammara 2008)	Yes	No	No	Yes	No	No	91.5%
(Meesad et al 2011)	No	No	Yes	Yes	No	Yes	92.20%
(Quan 2010)	No	No	Yes	Yes	No	No	90%
(Mitra et al. 2005)	Yes	No	No	Yes	No	No	99.66%
(Lee et al. 2010)	No	No	No	Yes	Yes	No	97%
(Remeikis et al. 2005)	Yes	No	Yes	No	No	No	90.9%

2.5. Text Mining in Content-based SNSs Using Clustering

Document clustering includes specific techniques and algorithms based on unsupervised document management [2.42]. In clustering the numbers, properties, and memberships of the classes are not known in advance [2.32]. Documents can be grouped together based on a specific category, such as medical, financial, and legal.

In scientific literature [2.25], [2.42], different clustering techniques are comprised of different strategies for identifying similar groups in the data. The clustering techniques can be divided into three broad categories: (a) hierarchical clustering, (b) partitional clustering, and (c) semantic based clustering that are detailed in the subsequent text.

2.5.1. Hierarchical Clustering

Hierarchical clustering organizes the group of documents into a tree like structure (dendrogram) where parent/child relationships can be viewed as a topic/subtopic relationship [2.43] Hierar- chical clustering can be performed either by using: (a) agglomerative or (b) divisive methods, which are detailed in the subsequent text [2.44].

An agglomerative method uses a bottom up approach by successively combining closest pairs of clusters together until the entire objects form one large cluster [2.44]. The closest cluster can be determined by calculating the distance between the objects of n dimensional space. Agglomerative algorithms are generally classified on the basis of inter-cluster similarity measurements. The most popular inter-cluster similarity measures are single-link, complete-link, and average-link [2.42]. Several algorithms are proposed based on the above mentioned approach [2.43], such as Slink, Clink, and Voortices use single-link, complete-link, and average-link, respectively. The Ward algorithm [2.43] uses both the agglomerative as well as divisive approach as illustrated in Figure 5. The only difference between the aforementioned algorithms is the method of computing the similarity between the clusters.

In [2.45], the authors suggested agglomerative hierarchal clustering techniques for text clustering. First, genetic algorithm was applied to achieve the feature selection phase in the text document. Second, similar document sets were grouped together into small clusters. Finally, the authors proposed text clustering algorithm to merge all clusters into final text cluster [2.45]. The proposed approach can be used for grouping the similar text from social networking websites, such as blogs, communities, and social media.

The divisive method uses a top-down approach by starting with the same cluster and recursively splitting the cluster into smaller clusters until each document is in a classified cluster [2.42]. The computations required by divisive clustering are more complex as compared to the agglomerative method. Therefore, the agglomerative approach is the more commonly used methodology.

Hierarchical clustering is very useful because of the structural hierarchal format. However the approach may suffer from a poor performance adjustment once the merge or split

operations are performed that generally leads to lower clustering accuracy [2.42]. Moreover, the clustering approach is not reversible and the derived results can be influenced by noise.

2.5.2. Partitional Clustering

Partitional clusters are also known as non-hierarchal clusters [2.44] To determine the relationship between objects, partitional clustering uses a feature vector matrix. Features of every object are compared and objects comprised of similar patterns are placed in a cluster [2.22]. The partitional clustering can be further categorized as iterative partitional clustering, where the algorithm repeats itself until a member object of the cluster stabilizes and becomes constant throughout the iterations. However, the number of clusters should be defined in advance [2.22]. Different forms of the iterative partitional cluster-based approaches are described as follows:

2.5.2.1. K-mean, K-medoid, C-mean, and C-medoid

In the k-mean approach the data set is divided into k clusters [2.42]. Each cluster can be represented by the mean of points termed as the centroid. The algorithm performs in a two-step iterative process: (1) assign all the points to the nearest centroid and (2) calculate the centroids for a newly updated group [2.42]. The iterative process continues until the cluster centroid becomes stabilized and remains constant [2.22].

The k-mean algorithm is widely used because of the straightforward parallelization [2.42]. Moreover, k-mean algorithm is insensitive to data ordering and works conveniently only with numerical attributes. However, the optimum value of k needs to be defined in advance [2.22].

The k-medoid algorithm selects the object closest to the center of the cluster to represent the cluster [2.42]. In the algorithm, the k object is selected randomly. Based on the selected object, distance is computed. The nearest object with respect to k will form a cluster. Remaining

objects take the place of k recursively until the quality of the cluster is improved [2.22]. The k-medoid algorithm has many improved versions, such as PAM (Partitioning around Medoid), CLARA (Clustering Large Applications), and CLARANS (Clustering Large Applications based upon Randomized Search). K-medoid algorithms work well for small data sets, but give compromised results for large data sets [2.22].

C-mean is a variation of k-mean that exhibits a fuzzy clustering concept that generates a given number of clusters with fuzzy boundaries and allows overlapping of clusters [2.43]. In overlapping clusters process, the boundaries of clusters are not clearly specified. Therefore, each object belongs to more than one cluster. Fuzzy C-Mean (FCM) [2.43] and Fuzzy C-Medoids (FCMdd) (Hang et al. 2008) algorithms are widely used examples of C-mean algorithm [2.43] as illustrated in Figure 5.

2.5.2.2. Single-pass Algorithm

The single-pass algorithm is the simplest form of partitional clustering [2.46] . The algorithm starts with empty clusters and randomly selects a document as a new cluster with only one member [2.46]. Single-pass algorithm calculates a similarity coefficient by considering a second object. If the calculated similarity coefficient is greater than the specified threshold value, then the object will be added to the existing cluster otherwise a new cluster will be created for the object. The BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm is an example of the single pass clustering algorithm [2.42]. The algorithm uses hierarchical data structure called CF tree for partitioning the datasets [2.46]. Nearest neighbour clustering is iterative and similar to the hierarchical single-link method [2.46].

2.5.2.3. Probabilistic Algorithm

Probabilistic clustering is an iterative method that calculates and assigns probabilities for the membership of an object [2.42]. Based on the probability measurements, an object can be a part of any specific cluster. Probabilistic clustering technique is popular because of the ability to handle records of a complex structure in a flexible manner. As probabilistic clustering has clear probabilistic foundations, finding out the most suitable number of clusters becomes relatively easy [2.22]. Examples of probabilistic clustering are the Exception Maximizing Algorithm (EMA) and Multiple Cause Mixture Model (MCMM). However, these approaches are computationally expensive [2.42].

2.5.3. Semantic-based Clustering

Meaningful sentences are composed of logical connections to meaningful words [2.22]. A logical construction of words is generally provided by machine readable dictionaries, such as WordNet. In semantic-based clustering, the structured patterns are extracted from an unstructured natural language. Moreover, the approach emphasizes meaningful analysis of contents for information retrieval.

Researchers have proposed several algorithms for computing semantic similarities between text, such as Resnick and Lin algorithms [2.22] are proposed to measure the semantic similarity of text in a specific taxonomy. Detailed descriptions of these algorithms are presented in [2.43].

Ref. [2.47] introduced a novel approach to automate the ontology construction process based on data clustering and pattern tree mining. The study comprises of two phases: (1) document clustering phase creates a group of related documents using k- mean clustering technique and (2) ontology construction phase creates inter-concept relation from the clustered

documents, whereas inter-concept relation is termed as similar concept relationship. The author implemented the proposed approach on weather news collected form e-paper and revealed remarkable results by extracting the regions with high temperature.

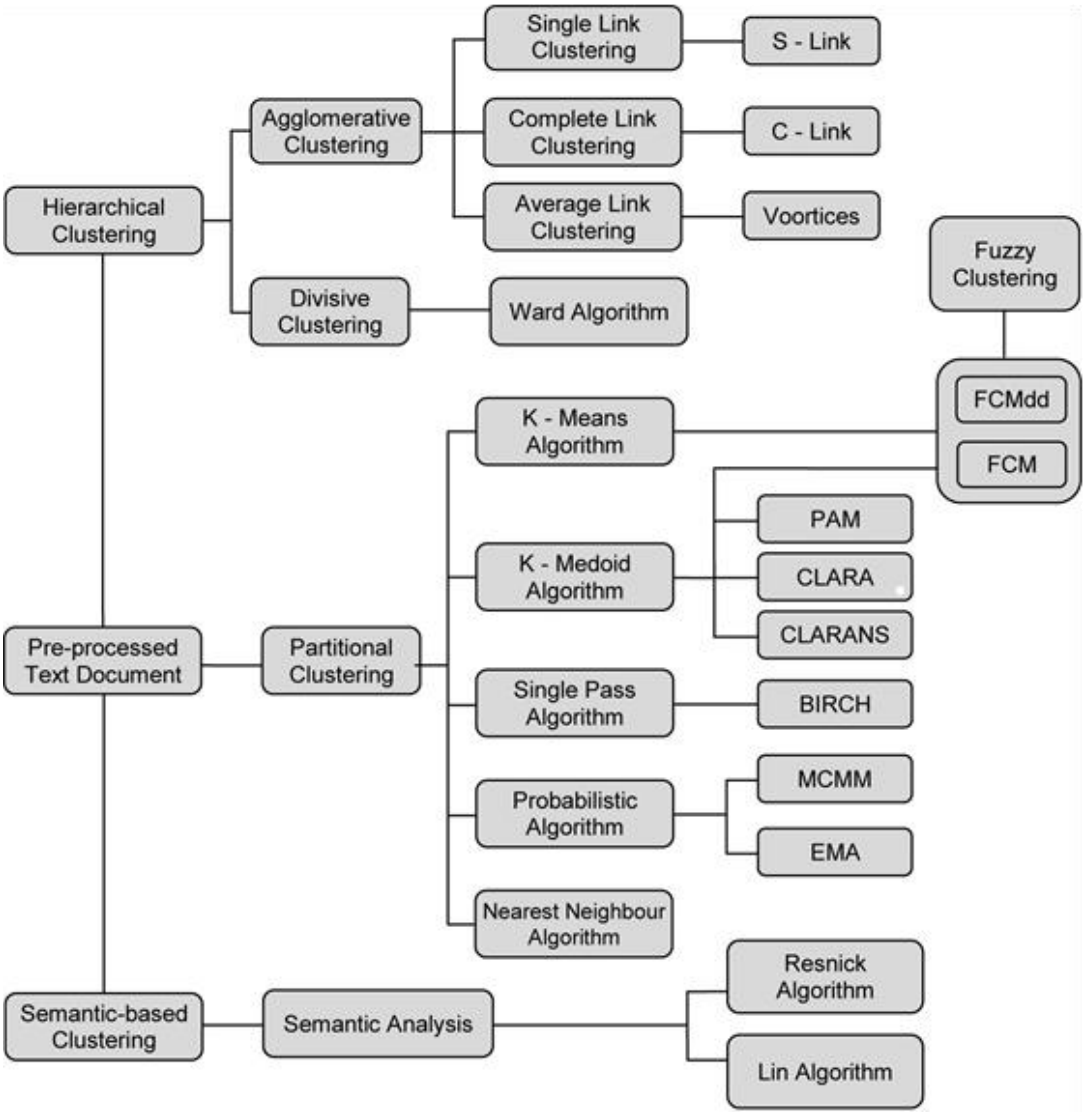

Figure 5. Text mining using clustering

2.6. Media-based Social Networking Services

A media-based SNS establish a social connection among users through the various multimedia formats, such as video and audio. Unlike content-based SNSs, media-based SNSs

have a high level of interactivity [2.48]. Different types of media-based-SNSs are characterized in the following subsections.

In recent years, we have witnessed a rapid growth of media-based SNSs, designed for providing interactions between the users within communities by sharing multimedia streams, such as photos and videos [2.48], [2.49]. The most popular SNPs that use photo/video sharing SNSs are Flicker, Instagram, and YouTube [2.48]. The photo/video sharing SNSs are popular due to their vast array of functions, such as: (a) real-time photo exchange (b) tagging, to describe the contents of an image (c) frame, that allows users to interpret a portion of an image (d) favorites, that allow users to select the most memorable collection from the group of images, and (e) comments, that allow annotation of the image in an appropriate manner [2.48], [2.3].

The photo/video sharing SNSs have been used extensively in business. Nov et al. in [2.48] incorporated photo/video sharing SNSs in the Integrated Marketing Communications (IMC) model to promote online communication between various companies and the customers. The use of photo/video sharing SNSs has dramatically improved the way consumers receive and react to the market information [2.48].

2.7. Geo-location-based Social Networking

The location-based SNSs are gaining popularity due to the advanced location-oriented hardware and software technologies, such as GPS-enabled devices, wireless communication technologies, and Internet connectivity through WiFi [2.6].

A holistic location-based social networking system termed as “GeoSocialDB” is introduced by Counts and Marc that provides the following location-based social services: (a) location-based news feeds (b) location-based news ranking, and (c) location-based recommendation [2.50] Moreover, Zheng et al. [2.6] specify three major research issues that

need to be addressed in the scalable implementation of location-based SNSs, namely: (a) designing a location-based query operation for the optimized query performance, (b) designing privacy-aware queries to protect the user location privacy, and (c) utilizing materialization techniques to accelerate the performance in terms of computation overhead and query response time [2.50], [2.6].

In real life scenario, people usually plan to visit places of interest while traveling to an unfamiliar location. However, proper travelling plans are not known in advance. To solve the aforementioned problem, GPS-trajectory-sharing presents an interactive approach to represent user's travel experiences and can provide reference for other users during the travel planning process to the unknown places [2.6], [2.5]. One's visited location histories can be tracked with a sequence of time-stamped locations, called trajectories [2.7]. The trajectories physically connect the visited locations in the world and provide information that can be further used for the experience sharing and geo-tagging multimedia content [2.7], [2.8], [2. 6]. In the scientific literature [2.50],[2. 5], various location-based SNSs have been discussed, such as GeoLife [2.7]. The GeoLife service performs three basic operations: (a) shares the life experiences, (b) provides the travel recommendations, such as top interesting locations, and (c) provides the friend recommendations based on the similarities among the location histories. However, an efficient approach is required to retrieve the user's desired GPS trajectory from a large-scale accumulated GPS dataset.

The most popular mobile devices, such as smart phones and location based mobile social networks, such as FourSquare [2.51], SoLoMo (Social Local Mobile) [2.52], and BrightKite [2.53] are gaining interest of the researchers, where the users can share the location with friends using the friendship networks. Based on the scientific literature reviewed, we observed that for

the accurate implementation of location based SNSs, there are two key challenges: (a) up-to-date information about the venues, such as nearby restaurant management, menu, food prices, and quality of food and (b) popularity ranking of the venue [2.6]. We believe that the above mentioned challenges can be address by introducing physical and logical contextual information in the current Geo-location-based SNSs.

Most of the content-based SNSs, such as Short Messaging Services (SMS), chatting, blogs, and Wikipedia were introduced in the 1980's and 1990's. The media-based SNSs, such as virtual world and video sharing were launched in 2002 and 2005 respectively. Moreover, the most promising geo-location-based SNSs were introduced in 2008 [2.9]. Furthermore, context-aware SNSs, such as social search and recommendations, have been attracting users' attention since 2009 through providing on-demand services to the users. A temporal timeline of the evolution of all the aforementioned SNSs are presented in Figure 4

Table 2 provides an exclusive overview of the main features and implication of different content-based, media-based, geo-location-based, and context-based SNSs in the various popular SNPs, such as Facebook, Flicker, and LinkedIn. It can be observed that the latest trend is to include the context-based technique into existing SNSs for better, real time, and on-demand communication. We hope that the presented issues will lead the researcher to explore the important research areas, such as on-demand collaboration, on-demand communication, social search, and context-aware recommendation systems.

Table 2. Social networking services in different social networking platforms

Social Networking Platforms (SNSs)	Content-based SNSs				Media-based SNSs		Geo-location-based SNSs	Context-based SNSs		
	Community blogs	Social News	Tagging	Chat	Audio / Video	Virtual World		Semantic Web	Social Search	Social Recommendation
Facebook	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Flicker	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
LinkedIn	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No
My Space	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Google +	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ipemity	Yes	No	Yes	Yes	Yes	No	Yes	No	No	No
YouTube	No	No	Yes	No	Yes	No	Yes	No	No	No
Orkut	Yes	No	Yes	Yes	Yes	No	No	No	No	No
Get Glue	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes
Live Joumal	Yes	No	Yes	No	Yes	Yes	No	Yes	No	Yes
Film Trust	No	No	Yes	No	No	No	No	Yes	No	Yes
Foursquare	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes

2.8. References

[2.1] Anderson, J., Diaz, C., Bonnean, j., Stajano. F.: 'Privacy-enabling social Netwrking over Untrusted Networks', Proc. Int. Conf. Online Social Networks, 2009, pp. 1-6

[2.2] Lops, p., Gemmis, M., and Semeraro, G.: 'Content Based Recommender Systems: State of the Art and Trends', 'Recommender System Handbook', R. Francesco, and R. Lior (Eds), Springer, 2011, pp. 73-105

[2.3] Lin, W.D., Jhong, S.Y., Huang, W., Lin, C.C.: 'Drawing Social Networks Using Area-Labeling Rectangular Cartograms', Journal of Internet Technology, 13(2), P.327-336 (2012/3) [2.4] D. Hirsch and S. Madria, "Data Replication in Cooperative Mobile Ad-Hoc Networks," Mobile Networks and Applications, vol. 18, no. 2, pp. 237-252, 2013.

[2.4] Sieg, A., Mobasher, B., and Burke, R.: 'Improving the Effectiveness of Collaborative Recommendation with Ontology-based User Profile'. Proc. Int. Conf. Information Heterogeneity and Fusion in Recommendation System, 2010, pp. 39-46

[2.5] Semeraro, G., Lops, P., Basile, P., Gemmis, M.: 'Knowledge Infusion into Content-based Recommender Systems'. Proc. Int. ACM Conf. Recommender Systems (RecSys), 2009, pp. 301-304

[2.6] Resnick, P., Varian, H.R.: 'Recommender Systems', Communication of the ACM, Mar. 1997, 40, (3), pp. 56-58

[2.7] Shankar, P., Huang, Y.W., H., Paul, C., Badri, N, and Liviu, I.: 'Crowds Replace Experts: Building better Location-based Services using Mobile Social Network Interactions'. Proc. Int. Conf. Pervasive Computing and Communication, 2012, pp.20-29

[2.8] Liu, F., Lee, H.J.: 'Use of Social Network Information to Enhance Collaborative Filtering Performance', Expert System with Applications, July 2010, 37, (7), pp. 4772-4778

[2.9] Sinha, R., and Swearingen, K.: 'Comparing Recommendation Made by Online Systems and Friends'. Proc. Workshop Personalization and Recommender Systems in Digital Libraries, 2001, pp. 45-56

[2.10] Karatzoqlou, A., Amatriain, X., Baltrunas, L., Oliver, N.: 'Multiverse recommendation: N-dimensional Tensor Factorization for Context-aware Collaborative Filtering'. Proc. Int. ACM Conf. Recommender Systems, 2010, pp. 79-86

[2.11] Panniello, U., Tuzhilin, A., Gorgoglione, M., Palmisano, C., and Pedone, A.: 'Experimental Comparison of Pre- vs Post-Filtering approaches in the Context Aware recommender System'. Proc. Int. ACM Conf. Recommender Systems, 2012, pp.265-268

[2.12] Sieg, A., Mobasher, B., and Burke, R.: 'Improving the Effectiveness of Collaborative Recommendation with Ontology-based User Profile'. Proc. Int. Conf. Information Heterogeneity and Fusion in Recommendation System, 2010, pp. 39-46

[2.13] Sieg, A., Mobasher, B., and Burke, R..: 'Web Search Personalization with Ontological user Profiles'. Proc. Int. ACM Conf. Information and Knowledge Management (CIKM), Nov. 2007, pp. 525-534

[2.14] Wang, Y., Wang, S., Stash, N., Aroyo, L., and Schreiber, G.: 'Enhancing Content-Based Recommendation with the Task Model of Classification'. Proc. Int. Conf. Knowledge and Management, 2010, pp. 431-440

[2.15] Irfan, R., King, C., Grages, D., Ewen, S., Khan, S.U., Madani, S.A., J. Kolodziej, Wang, L., Chen, D., and Rayes, A.: 'A Survey on Text Mining in Social Networks', Knowledge Engineering Review

[2.16] Y. Doytsher, B. Galon, and Y. Kanza, "Storing routes in socio-spatial networks and supporting social-based route recommendation," In Proc. 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, ACM, pp. 49-56, 2011.

[2.17] Y. Zheng, L. Zhang, X. Xie, and W.Y. Ma, “Mining interesting locations and travel sequences from GPS trajectories,” In Proceedings of the 18th international conference on World wide web, ACM, pp. 791-800, 2009.

[2.18] L. Wei, Y. Zheng, and W. Peng, “Constructing popular routes from uncertain trajectories,” In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 195-203, 2012.

[2.19] Batet, M., Valls, A., Gibert, K., and Sanchez, D.: ‘Semantic Clustering Using Multiple Ontologies’. Proc. Int. Conf. Catalan Association for Artificial Intelligence, 2010, pp. 207-216

[2.20] Y Kulkarni, D., and Tripathi, A.: ‘A Framework for Programming Robust Context-Aware Applications’, IEEE Transaction on Software Engineering, April 2010, 36, (2), pp. 184-197

[2.21] Sorensen, L. 2009. User managed trust in social networking comparing facebook, myspace and linkdin. In Proceed- ings of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic System Technology, (Wireless VITAE 09), Denmark, 427-431.

[2.22] Liu, F. & Lu, X. 2011. Survey on text clustering algorithm. In Proceedings of 2nd International IEEE Conference on Software Engineering and Services Science (ICSESS), China, 901-904.

[2.23] Kano,Y., Baumgartner, W. A., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L. & Tsujii, T. 2009. Data Mining: Concept and Techniques. Oxford Journal of Bioinformatics, 25(15), 1997-1998

[2.24] Yin, S., Wang, G., Qiu, Y. & Zhang, W. 2007. Research and implement of classification algorithm on web text mining.In Proceedings of 3rd International Conference on Semantics, Knowledge and Grid, China, 446-449.

[2.25] Xu, X., Zhang, F. & Niu, Z. 2008. An ontology-based query system for digital libraries. In Proceedings of IEEE, Pacific-Asia Workshop on Computational Intelligence and Industrial Application, Wuhan, 222-226.

[2.26] Chang, M. & Poon, C. K. 2009. Using phrases as features in e-mail classification. Journal of System and Software, 82(6), 1036-1945.

[2.27] Forman, G. & Kirshenbaum, E. 2008. Extremely fast text feature extraction for classification and indexing. In Proceedings of 17th ACM Conference on Information and Knowledge Management, California, USA, 26-30. Gazzah,

[2.28] Patterson, D., Rooney, N., Galushka, M., Dobrynin, V. & Smirnova, E. 2008. SOPHIA-TCBR: A knowledge discovery framework for textual case-based reasoning. Knowledge-Based Systems, 21(5), 404-414.

[2.29] Baumer, E. P. S., Sinclair, J. & Tomlinson, B. 2010. America is like metamucil: Fostering critical and creative thinking about metaphor in political blogs. In Proceedings of 28th International Conference on Human Factor in Computing Systems (CHI 2010) ACM, Atlanta, GA, USA, 34-45.

[2.30] Jo, T. 2010. NTC (Neural Text Categorizer): Neural network for text categorization. International Journal of Information Science, 2(2), 83-96.

[2.31] Kolodziej, J., Burczynski, B. & Khan, S. U. 2012. Advances in Intelligent Modelling and Simulation: Artificial Intelligence-Based Models and Techniques in Scalable Computing, New York. Springer-Verlag.

[2.32] Luger, G. F. 2008. Artificial Intelligence: Structure and Strategies for Complex Problem Solving. 6th edn. Addison Wesley.

[2.33] Kolodziej, J. & Khan, S. U. & Xhafa, F. 2011. Genetic algorithms for energy-aware scheduling in computational grids. In Proceedings of 6th IEEE International Conference on P2P, Parallel, Grid, Cloud, and Internet Computing (3PGCIC), Barcelona, Spain, 17-24.

[2.34] Guzek, M., Pecero, J. E., Dorronsoro, B., Bouvry, P. & Khan, S. U. 2010. A cellular genetic algorithm for scheduling applications and energy-aware communication optimization. In Proceedings of PACM/IEEE/IFIP International Conference on High Performance Computing and Simulation (HPCS), Caen, France, 241-248.

[2.35] Zhao, Y. & Dong, J. 2009. Ontology classification for semantic-web-based software engineering. IEEE Transactions on Service Computing, 2(4), 303-317.

[2.36] Li, J., Wang, H. & Khan, S. U. 2012. A fully distributed Scheme for discovery of semantic relationships. IEEE Transactions on Services Computing.

[2.37] Wimalasuriya, D. C. & Dou, D. 2010. Ontology-based information extraction: An introduction and a survey of current approach. Journal of Information Science, 36(5), 306-323.

[2.38] Miao, D., Duan, Q., Zhang, H. & Jiao, N. 2009. Rough set based hybrid algorithm for text classification. Journal of Expert Systems with Applications, 36(5), 9168-9174.

[2.39] Aci, M., Inan, C. & Avci, M. 2010. A hybrid classification method of k-nearest neighbour, bayesian method and genetic algorithm. Expert Systems with Applications, 37(7), 5061-5067.

[2.40] Meesad, P., Boonrawd, P. & Nuipian, V. 2011. A Chi-square-test for word importance differentiation in text classification. In Proceedings of International Conference on Information and Electronics Engineering, Singapore, 110-114.

[2.41] Li, J., Li, Q., Khan, S. U. & Ghani, N. 2011. Community-based cloud for emergency management. In Proceedings of the 6th IEEE International Conference on System of Systems Engineering (SoSE), Albuquerque, USA, 55-60.

[2.42] Jain, A. K. 2010. Data clustering: 50 years beyond k-means. Pattern Recognition, 31(8), 651-666.

[2.43] Chen, W. & Wang, M. 2009. A fuzzy c-means clustering-based fragile watermarking scheme for image authentication. Expert Systems with Applications, 36(2), 1300-1307.

[2.44] Kavitha, V. & Punithavalli, M. 2010. Clustering time series data stream - A literature survey. International Journal of Computer Science and Information Security, 8(1), 289-294.

[2.45] Yonghong, Y. & Wenyang, B. 2010. Text clustering based on term weights automatic partition. In Proceedings of 2nd International Conference on Computer and Automation Engineering (ICCAE), China, 373-377.

[2.46] Mehmed, K. 2011. Data Mining: Concepts, Models, Methods, and Algorithms, 2nd edn, Canada, John Willy & Sons.

[2.47] Yu, Y. & Hsu, C. 2011. A structured ontology construction by using data clustering and parttern tree mining. In Proceedings of International Conference on Machine Learning and Cybernetics, Guilin, 45-49.

[2.48] Nov, O., Naaman, M., Ye, C.: 'Analysis of Participation in an Online Photo-Sharing Community: A Multidimensional Perspective', Journal of American Society for Information Science and Technology, May 2010, 61, (3), pp. 555-556

[2.49] Lerman, K., Jones, L.: 'Social Browsing on Flickr'. Proc. Int. Conf. Weblogs and Social Media (ICWSM), 2007, pp. 125-138

[2.50] Counts, S., Smith, M.: 'Where Were We: Communities for Sharing Space-time Trails'. Proc. Int. Conf. Advance in Geographic Information Systems, 2007, pp. 127-135.

[2.51] https://foursquare.com/, Foursquare, Accessed 17th January 2013.

[2.52] http://techcrunch.com/2012/10/11/solomo-update-mobile-only-social-networks-to-reach-1b-users-by-2014-800m-users-of-foursquare-other-location-services-this-year/, SOLOMO, accessed 17th January2013.

[2.53] http://en.wikipedia.org/wiki/Brightkite, Brightkite, Accessed 17th January 2013.

3. ONTOLOGY LEARNING IN TEXT MINING

3.1. Abstract

The dynamic visualization of text due to the current Read-Write-Web (RWW) provides an enormous and growing source of information. However, extracting required information and sharing it in different application remain a challenging task. The categorization of unstructured text is one of the fundamental data analysis techniques that have been widely studied in the various disciplines for indexing, mining, and managing abundant textual data. Ontology offers the potential for providing a logical interpretation of textual data that is based on a hierarchical conceptual representation of information. However, one of the major obstacles that prevents ontology from being deployed in large-scale information systems is ontology acquisition, which strongly depends on knowledge engineers and domain experts. Additionally, ontology building is a labor-intensive, handcrafted, and recursive process. Therefore, to address the abovementioned problem, researchers have devised semi-automatic techniques called ontology learning for building ontologies. This survey provides a comprehensive analysis of ontology learning techniques, such as linguistic, statistical, and semantic-based techniques, extensively used in ontology learning. Moreover, the survey provides a detailed review of the ontology learning process. The discussion moves on further to present the ontology-based text mining architecture and highlights various attempts of scientific researchers to successfully incorporate ontologies in the field of text mining. Furthermore, we identify major issues and challenges in the ontology learning process that need to be addressed in future semantic-based text extraction efforts.

Keywords: Ontology, text mining, ontology learning process, ontology learning techniques, categorization, classification, linguistics, semantic, term, axiom.

3.2. Introduction

Ontologies enable knowledge sharing and provide a logical and machine-interpretable format for the unstructured textual data. The term ontology mostly refers to a semantic container that provides common vocabularies (meta-data) associated with a specific domain [3.1], [3.2], [3.3]. Gruber defines ontologies as "explicit formal specification of a shared conceptualization" [3.6], where explicit means that the type of concepts and the constraints on their use are explicitly defined, formal means that the ontology should be machine readable, shared reflects the notion that an ontology acquires consensual knowledge, and conceptualization emphasizes the abstract model of some phenomenon in the world through classifying the relevant concept. The study of ontologies has developed gradually from specific needs associated with the problem of knowledge management within a computational environment and particularly from the problem of knowledge sharing and reuse [3.6]. Ontologies overcome the difficulties raised by monolithic, isolated knowledge systems by specifying a content-specific agreement to facilitate the knowledge sharing and reuse in a specific domain [3.47]. Moreover, ontologies comprise: (i) objects, (ii) concepts, and (iii) hierarchical relationships between concepts and objects to acquire semantics in a specific field, such as medicine, academia, and engineering [3.42], [3.27], [3.31], [3.2], [3.10], [3.23], [3.49]. For instance, medical ontology contains the basic concepts related to treatments of various diseases and clinical procedures that facilitate the propagation of standard medical terminology in the healthcare systems. Currently, the most popular ontologies are Protein Ontology (PO), Basic Formal Ontology (BFO), Unified Medical Language system (UMLS), Suggested Upper Merged Ontology (SUMO), and Bio Investigation Ontology (BIO) [3.23], [3.27], [3.38].

The concept of ontology has been instrumental in many other application areas, such as semantic search, semantic Web, and text mining, because of the flexible annotation capabilities that are acquired through the formation of hierarchical concepts [3.5], [3.3]. Various text mining techniques, such as classification and clustering, statistically evaluate the occurrence of the words and group similar documents together in the text corpora [3.4], [3.5], [3.3]. In the scientific literature, numerous methods have been investigated and applications implemented in the area of text mining, ranging from hierarchical categorization of Web pages to automated text generation. However, the majority of the text mining techniques incorporate traditional statistical-based approaches using machine learning and pattern recognition methods [3.41], [3.15], [3.18]. The ontologies are used mainly as background knowledge for representing concept hierarchy and for providing semantics to the textual document [3.32], [3.9], [3.13], [3.43]. Nevertheless, the gains from using ontologies in the field of text mining have become clear in recent years [3.27], [3.37], [3.38], [3.42], [3.45]. Scientific literature shows numerous attempts to integrate ontology into the field of text mining [3.31], [3.9].

Indeed, ontology is the backbone of future text-based information extraction systems. However, one of the major obstacles that prevent ontology from being deployed in large-scale information systems is the manual knowledge acquisition process that strongly depends on knowledge engineers and domain experts [3.31]. Building the ontology manually presents a major knowledge acquisition bottleneck for several reasons, such as: (i) dynamic expansion of text, (ii) unavailability of human experts and knowledge engineers, and (iii) rapid and distributed evolvement of domain knowledge [3.12], [3.40], [3.42], [3.47]. Moreover, groups of domain specialists and knowledge engineers must work continuously to keep the ontology up-to-date [3.41]. Therefore, the manual ontology building process is time-consuming, labor-intensive,

expensive, and error-prone [3.41], [3.31]. In recent years, researchers have attempted to automate the ontology building process, called ontology learning. However, most of the scientific literature on ontology learning does not provide an inclusive analysis of various ontology learning techniques, such as linguistic, statistical, and semantic, for the ontology learning process [3.12[, [3.13], [3.41], [3.40], [3.42], [3.38], [3.83]. Linguistic-based techniques are language-dependent and used exclusively for the syntactical analysis of the unstructured text [3.41]. Statistical techniques provide a statistical analysis of the text corpora. Most of the statistical-based techniques are derived from the field of machine learning and text mining [3.22]. Semantic-based techniques use a logical interpretation of the textual contents. Most of the scientific literature on ontology learning does not provide an inclusive analysis of such techniques. Moreover, existing surveys [3.12], [3.7], [3.41] do not discuss structured and domain-specific ontology learning implications.

Our survey focuses on three major aspects: (i) ontology learning techniques, (ii) the ontology learning process, and (iii) the implication of various ontology learning techniques in the ontology learning process. Moreover, the discussion presents the ontology-based text mining architecture as an example of ontology learning implication into the field of text mining, the most promising research area for logical interpretation of the text corpora. Furthermore, we review the major issues and challenges in the ontology learning process. The paper sheds light on the latest research about ontologies and paves the way for future research in the ontology learning process.

The rest of the paper is organized as follows. Section 2 presents a comprehensive study on ontology learning from unstructured text and explains the ontology learning process with a detailed description of ontology learning techniques. The ontology-based text mining

architecture is described in Section 3. Section 4 discusses various current advancements in the field of ontology learning process. Section 5 reviews various open questions with brief discussions on current issues of the ontology learning process in text mining. Finally, Section 6 concludes the survey with three important direction for future research.

3.3. Ontology Learning from Text

Because of the pervasive use of ontologies for the logical interpretation of unstructured text, researchers from various fields, such as knowledge engineering, computational linguistics, information retrieval, and text mining have acknowledged the need for using effective techniques to automate the ontology building process [3.3], [3.47, [3.67]. Manual ontology construction demands considerable time and effort from domain experts and ontology engineers [3.3]. Ontology learning is a semi-automatic process that requires human validation and consensus to represent the conceptualization of a specific domain for the ontology building process [3.3], [3.4]. A detailed description of ontology learning process and ontology learning technique is presented in the subsequent text.

3.3.1. Ontology Learning Process

An ontology learning process comprises five modules: (i) terms, (ii) synonyms, (iii) concepts, (iv) relationships, and (v) axioms or rules [3.23], [3.15], [3.13], [3.21]. The detailed description of the modules is presented in the subsequent text.

a. A term is a linguistic realization of the specific concept and a basic unit of the ontology learning process [3.15]. Preprocessing and term extraction are two tasks associated with a term. Preprocessing truncates the term into a format that ensures the successful implication of the ontology learning process [3.1]. Three basic techniques of pre-processing are: (i)

tokenization, (ii) stop-word-removal, and (iii) word stemming, as described in following Section 2.2 (2.1.1).

Term extraction extracts the relevant words from a document [3.23], [3.1]. The extraction phase comprises two steps: (i) counting the frequency of terms and (ii) extracting the concept with similar context [3.18].

b. Synonyms describe the semantic similarities between different words [3.15]. Semantic similarity measures the degree to which two different terms resemble each other, whether in a single language or in multiple languages [3.13]. The WordNet is a popular lexical database that groups the words together, based on similar meanings EuroWordNet is a lexical database that is used for the bilingual and multilingual synonyms extraction [3.22], [3.13], [3.15]. Other examples of synonyms are WordNet-Affect and SentiWordNet, which represent the words that possess similar emotion [3.13], [3.67]. In addition to using built-in synsets provided by WordNet and EuroWordNet, researchers have exploited various word-sense disambiguation algorithms for synonyms acquisition [3.21], [3.13]. Most of the algorithms are based on the Harris distributional hypothesis, according to which the semantic similarity of the words can be measured by the extent to which the words share the syntactic and semantic context [3.21]. For instance, the word “mouse” stands for an animal when considered within the context of the animal kingdom and a peripheral device when considered in the context of computers and technologies. Additionally, various statistical-based techniques, such as Latent Semantic Indexing algorithms (LSI), Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Indexing (PLSI) and linguistic techniques, such as lexico-syntactic patterns and parsing are widely used to identifying the inherent connections between the terms [3.13], [3.21].

c. The concept module represents a group of similar items. Various lexical semantic categories from WordNet are used to map the extracted terms with the different concepts termed as concept mapping as presented in the Figure 4. Moreover, various sophisticated term disambiguation strategies, such as bootstrapping and Machine Readable Dictionaries. (MRDs), are used to extract the best sense combination, when multiword terms are involved in the logic building process [3.36], [3.14].

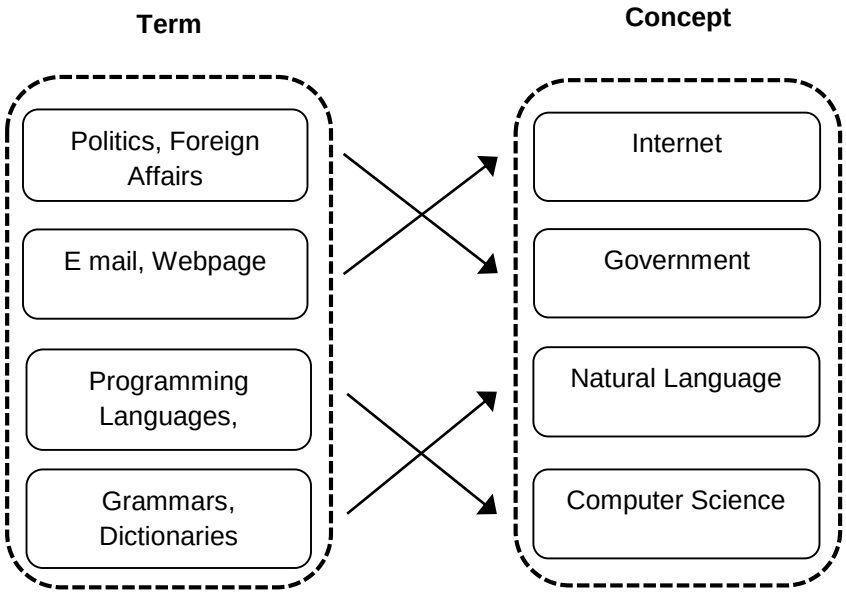

Figure 6. Concept mapping

d. The relationship refers to a process of modeling an association between different concepts [3.14]. There are two types of relations, taxonomic and non-taxonomic. A taxonomic relation is widely applicable to represent the generalized hierarchical relationships, such as “is-a” and “has-a,” and hyponymy, which presents simple and multiple inheritance between the words [3.14]. A non-taxonomic relation is the connection between the concepts, based on multi-word expressions, such as meronymy and antonymy [3.13], [3.3]. A non-taxonomic relation

discovery and labeling is more complex because of the explicit way of concept presentation [3.14]. For a detailed overview on extraction of taxonomic and non-taxonomic relations from the text, readers are encouraged to consult the work of Seera et al. [3.41].

Axiom building is the final step of the ontology learning process. An axiom is a preposition or a set of rules that are used for the evaluation of the specific domain [3.3]. Most axioms are formalized by using first-order logic and decision trees to encode different terminologies [3.20]. Moreover, axioms can be utilized to verify the correctness of existing ontologies [3.20], [3.62].

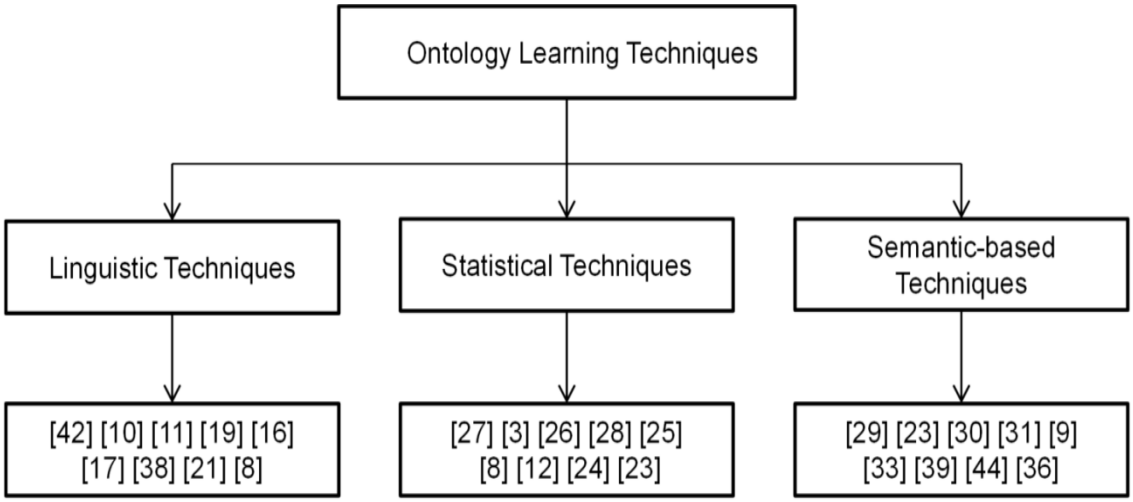

Figure 7. Taxonomy of ontology learning techniques

3.3.2. Ontology Learning Technique

Knowledge acquisition from unstructured text has become a promising field as a result of the explosion of the Read-Write-Web (RWW) [3.40]. Various techniques, such as text mining, Natural Language Processing (NLP), machine learning, Information Extraction (IE), and Artificial Intelligence (AI) have contributed enough to build semi-automatic methods for the ontology learning process. The text mining and machine learning techniques extract logical

patterns from the unstructured text corpora by using statistical techniques, such as Naive Bayes, K-means, and Support Vector Machine (SVM) [3.40]. The field of NLP provides tools for the analysis of text with reference to grammatical structure for concept identification. Moreover, AI provides logic-based inference techniques to deduce new concepts from the text.

Ontology learning techniques can be categorized as follows: (i) linguistic, (ii) statistical, and (iii) semantic-based [3.3], [3.12], [3.27]. Figure 7 presents the taxonomy of ontology learning techniques. Various ontology learning techniques that are applicable in the ontology learning process are presented in the following subsections.

3.3.2.1. Linguistic Techniques

Linguistics techniques are applicable to almost all modules of the ontology learning process. In the scientific literature, various linguistics techniques are used for analyzing the grammatical structure of a sentence. The grammatical structure refers to the building blocks of a language that governs the accurate composition of words in a sentence. Most of the linguistic techniques are language-dependent and are based on NLP tools [3.40]. Linguistic techniques comprise morphological techniques and syntactic techniques described as follows.

The morphological techniques are based on the extraction of the individual words in a sentence that contains information. For instance, tokenization is the process of splitting the sentence into different words, such as number, punctuation marks, and names [3.10]. JavaTok is a free configurable tokenizer, developed in Java [3.84] and Efficient Tokenizer (ET) is a popular tokenizer written in standard Prolog language [3.85]. Another morphological technique is remove-stop-word. Stop-words, such as “the,” “am,” “an,” and “a,” construct the syntactic structure of the sentence and are the most frequently occurring words. However, these words do not contribute enough to represent the information. Therefore, stop-words are removed from the

text corpus in the information extraction process. Stemming is another morphological technique that refers to a linguistic normalization to remove the prefixes and suffixes from a word. For instance, the word “connection” is reduced to the root word “connect.” The stemming technique has been widely used in various algorithms, such as n-gram analysis, stochastic algorithms, Affix stemmers, and poster stemmer [3.18], [3.10].

A syntactical analysis provides information about the grammatical structure of a sentence in the text document. Various techniques are available in the scientific literature for syntax analysis of the sentence, such as parsing, part-of-speech tagging, sub-categorization frame, syntactic structure, analysis/dependency analysis, WordNet dictionary, and lexico-syntactic [3.15], [3.16], [3.36], [3.20]. The detailed description of the aforementioned techniques is presented in subsequent text.

a. Parsing and part of speech tagging generate a hierarchical structure termed as parse-tree that presents a syntactic arrangement of the words in a sentence by using parts of speech, such as verbs, nouns, adjectives, and adverbs [3.20], [3.61]. The part of speech tagging is also termed as grammatical tagging that markup the words in a corpus analogous to a particular part of speech. Common examples of sentence parsers and part-of-speech taggers are Sage [3.15], Brill Tagger [3.38], Minipar [3.42], and Principle-based parser (Principar) [3.15]. Parsing and part of speech tagging is extensively used to extract terms at a sentence level in the ontology learning process.

b. Sub-Categorization Frame (SCF) and Syntactic Structure Analyzer (SSA) techniques are commonly used to create a syntactic structure of the sentence [3.16]. SCFs are considered a key component of a computational lexicon because of the effective word categorization based on the syntax in a text document. An SSA is an extended version of SCF [3.16]. In

SSA, words in syntactic structures, such as noun (NN), verb, adjectives (ADJ), and prepositional phrases are evaluated to discover potential terms and relations. For instance, ADJ-NN can be extracted as a potential term while overlooking other part of speech. The major difference between parsing and SSA is that the parsing provides syntactic arrangement of the words for further linguistic analysis. Whereas, SSA provides word categorization to identify relations between parts of speech. In ontology learning process, various SCFs and SSAs-based tools are available, such as the Standford NLP [3.36], and GATE (General Architecture for Text Engineering) [3.19] to determine concept and more complex relations (taxonomic, non-taxonomic) from the textual data.

c. WordNet is a semantic dictionary comprising an organized set of similar words (synonyms) in a form of unordered set termed as synset. The synset contains brief definitions of synonyms to illustrate the lexical relations between items with similar implication [3.36], [3.57], [3.67]. The WordNet is widely used in the area of word sense disambiguation to extract the exact meaning of a word by utilizing synset in a scenario where a single word has multiple meanings [3.8]. For instance, the word “pitcher” can be categorized as (pitcher, Jug) and (pitcher, baseball player) to disambiguate the synonyms. Recently, BabelNet a multilingual lexicalized ontology has been introduced that linked Wikipedia the largest Web-based encyclopedia to WordNet [3.68]. The BebelNet provides lexicographic and encyclopedic interpretation of terms by utilizing semantic networks for logical analysis of a term. Semantic networks connect the items in a large network of semantic relations that are derived from Babel synset. Each Babel synset presents a meaning and provides all possible synonyms of a term in a range of different languages [3.68].

All aforementioned techniques are used extensively in the ontology building process. However, these techniques produce a huge amount of lexical or grammatical data, such as "is," "have," and "a," that is trivial to compute [3.8]. Moreover, analyzing and updating the lexical semantic relationships between vocabularies, such as synonyms and context, are computationally expensive [3.8]. At present, the most popular and successful linguistic techniques for ontology building include parsing, WordNet dictionaries, and SSA [3.67], [3.51], [3.68], [3.8].

3.3.2.2. Statistical Techniques

Majority of statistical techniques for ontology learning process are derived from the field of data mining, information retrieval, and machine learning. Statistical techniques are used to evaluate the occurrence of words in a document and create mathematical models to derive logical patterns from the text corpora. An occurrence of a word in a textual corpus provides a reliable estimation of the semantics of a document [3.25]. The semantic identity of a term can be represented by analyzing the distribution of the words in the entire document [3.3]. Statistical techniques include clustering, classification, co-occurrence analysis, and Latent Semantic Analysis (LSA). The description of statistical techniques is as follows:

a. Clustering and classification are used to extract logical patterns from the textual data [3.24]. The term clustering refers to an unsupervised learning that categorizes the text into groups based on the common characteristics. Alternatively, classification is a supervised learning that precedes statistical evaluation of words from a training set and assigns the closest items or concepts into a group [3.25]. Recently, S. Ray et al. proposed an automated text classification technique that categorizes the unstructured text by utilizing statistical approach termed as Term Frequency- Inverse Document Frequency (TF-IDF) [3.62]. However the proposed technique considers the terminologies associated with a specific domain and does

not take into account the underlying meaning of the words used in a corpora. The most popular clustering techniques that are used in the ontology learning process are hierarchical clustering and the Naïve Bayesian (NB) approach [3.46], [3.58], [3.63]. The hierarchical clustering technique represents a relationship between topics and subtopics in a document as a tree-like structure termed as dendrogram [3.24], [3.25]. Hierarchical clustering uses two approaches: (i) agglomerative (bottom up) and (ii) divisive (top down). Agglomerative clustering is a commonly used clustering technique in the ontology learning process [3.24], [3.25]. In the scientific literature, several algorithms, such as Clink, Voortices, Ward, and Slink are proposed for the identification of concepts and relations in the ontology learning process [3.26], [3.35]. The NB approach provides a probabilistic approach to identify word occurrences and their associated concepts within a particular document [3.46], [3.24]. Moreover, in the ontology learning process, a decision tree is also a popular technique for classification to detect the generalized association between the items and concepts. Interested readers are encouraged to consult the Minipar website [3.42], and [3.8], [3.52], [3.53], [3.55] for a detailed description of classification and clustering techniques.

b. Latent Semantic Analysis (LSA) estimates the correlation between words at the conceptual level to discover the implicit relationships between concepts in an ontology learning process [3.23]. Similarly, co-occurrence analysis finds relationships among words and extracts the lexicon unit that tends to occur together [3.23], [3.12]. For instance, the phrasal level co-occurrence presents “United” and “States” as a single word depicting a single entity. Co-occurrence analysis also presents common logical associations between different words, such as “knife” and “cut.” The basic purpose of co-occurrence analysis is to extract related terms that play a key role to discover relations between concepts in the ontology learning process

[3.22]. Both of the aforementioned analyses use conditional probabilities to discover the hierarchical relationships between words and concepts. Some of the popular co-occurrence measure techniques are: (i) log-likelihood relation, such as chi-square [3.22], (ii) Term-Frequency-Inverse Document Frequency (TF-IDF) [3.23], (iii) Mutual Information Measure (MIM) [3.22], (iv) similarity measure [3.25], and (v) Kullback-Leiber divergence [3.22].

c. Association rule mining is another statistical technique extensively employed to discover taxonomic and non-taxonomic conceptual relation in the ontology learning process [3.73], [3.10]. Such technique presents associations between different concepts in a predefined level of abstraction [3.74], [3.73]. For instance, a very popular example of association rule mining is the market basket analysis, discussed by Srikant and Agrawal in [3.74]. The authors identified the products that customer frequently purchase together in a same transaction by utilizing association rule mining. In the past few years, the implication of association rule mining in ontology learning process is evident. Recently, Brisson et al. [3.73] proposed an Ontology Driven Information System (ODIS) that integrates prior knowledge about a specific domain with data mining process in a coherent and uniform manner. In the data mining process, the authors used association rule extraction algorithm that generates maximum non-redundant relationships between items in a form of generalized rules. Similarly, the author Antinio et al. in [3.75] proposed an association rule mining model that is explicitly tailored to support Ontological Knowledge Base (OKB). Association rule mining has the potential to discover frequent item sets and define additional probabilistic connections between these item sets that can be utilized as a basis of ontology learning process [3.75], [3.73].

All these discussed statistical techniques have been widely used in ontology building for the past few decades. However, the limitations of statistical techniques include the following: (i) lack of expressing assertions about real-life elements (propositional logic), (ii) lack of background knowledge (common sense) representation techniques, (iii) lack of effective generalization or pattern identification process, and (iv) lack of logical association extraction among words [3.21].

3.3.2.3. Semantic-based Techniques

The limitations of statistical techniques have been resolved by using semantic-based techniques. These techniques use induction and deduction-based reasoning to discover new facts from the existing knowledge. Such techniques are generally used in the concept, relation, and axiom module of ontology learning process. Inductive reasoning induces the hypothesis from a specific example and derives new knowledge from the presented text [3.27]. Deductive reasoning exploits inference rules, such as resolutions to deduce new knowledge. Various semantic-based techniques are available in the scientific literature to extract logics from text corpora, such as: (i) Inductive Logical Programming (ILP) and (ii) inference techniques [3.21]. Semantic-based techniques are presented in subsequent paragraphs.

a. An intersection of machine learning and inductive learning fields is known as Inductive Logical programming (ILP) [3.28]. ILP generates rules from observed instances and synthesizes new knowledge based on the past observations [3.21]. The most popular techniques for inductive logic programming are Winston's learning and version space [3.28]. Moreover, knowledge representation techniques, such as frame, scripts, and conceptual graphs use logical programming for the conceptualization of the background knowledge about a specific concept [3.22], [3.31]. Various rule-based induction and instance-based learning

algorithms, such as RISE, and First Order Inductive Learning (FOIL) are used for building the ontology learning process [3.28], [3.29], [3.9].

b. Inference is a powerful method to derive conclusions from the premises or hypothesis [3.22]. For instance, from the presence of smoke in the building one might infer the presence of a fire, or from the sentences "All basketball players are tall, and John is a basketball player", one can infer that "John is tall." The basic components of Inference include: (i) premises or input (ii) application of inductive or deductive reasoning, and (iii) generation of output or results. Various methods have been used to implement inferences, such as unification, generalized modus ponens, modus tollens, backward chaining, and forward chaining [3.30]. In text mining, however, the inference techniques and ILP adds computation complexity because of the unstructured format of textual data [3.28].

Figure 8 presents the association of ontology learning process with different ontology learning techniques. Recent studies also introduced a hybrid approach (linguistic-statistical-semantic techniques) called C-value/NC-value to extract single-word and multi-word terms from the textual data that can used further for the ontology learning process [3.15]

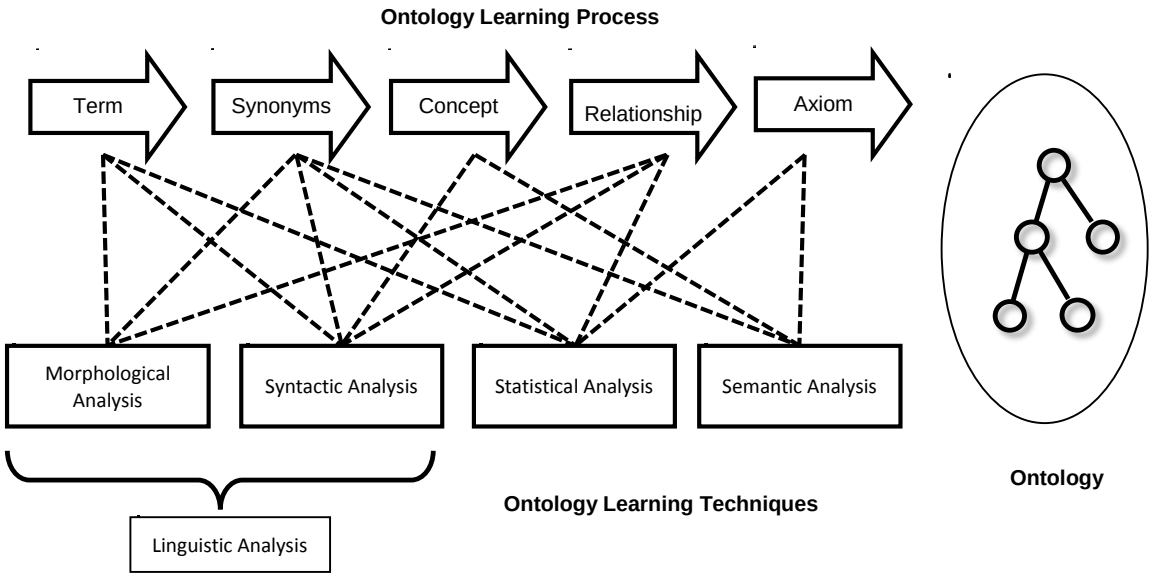

Figure 8. Association of ontology learning with ontology learning

In the scientific literature, various ontology learning techniques have been proposed to extract the relevant terms, synonyms, concepts, relationships, and axiom to automate the ontology acquisition process. Table 3 shows a coherent analysis of various ontology learning techniques applicable to different stages of the ontology learning process.

Table 3. Ontology learning techniques in the ontology learning process

Ontology Learning Process		Ontology Learning Techniques			
		Linguistic Technique		Statistical Technique	Semantic-based Technique
		Morphological Analysis	Syntactic Technique		
Term	Pre-processing	Tokenization, Remove stop-words, word Stemming	Affix-Removal, Suffix-Removal	N gram,	
	Term Extraction	Remove stop word	POS Tagging, Parsing, Sub categorization frame, Syntactic structure analysis/ dependency analysis	Co-occurrence analysis, Relevance analysis, Term weighting (TFIDF), Chi square, Mutual information, Hidden Markov Model (HMM), Lexical Probabilistic parser (Stanford parser), Latent semantic indexing	
Synonym	Word Identification	Word stemming	WordNet, Machine readable dictionaries	Latent semantic indexing	Semantic lexicon, Frame, Script
Concept	Concept Extraction	Word stemming	Sub categorization frame	Latent semantic indexing, Co-occurrence analysis, Agglomerative clustering	Semantic lexicon, Logical inference, Semantic template
	Concept Label			Clustering, Classification	
Relationship	Hierarchical Construction, Taxonomic Relations		Machine readable dictionaries, Syntactic structure/dependency analysis, Lexicon-patterns analysis	Co-occurrence analysis	Semantic lexicon, Scripts, Logical inference induction and deduction.
	Non-Taxonomic Relation Discovery	Tokenization, Remove stop words	WordNet, Syntactic structure/dependency analysis, Lexicon-pattern analysis.	Association rule mining, Decision tree	Logical inference, Frame-based language (OKBC)
	Non-Taxonomic Relation Labeling	Remove stop-word	Syntactic structure/Dependency analysis.	Conditional probability	Semantic lexicon
Axiom	Axiom Extraction			Inference based clustering	Logical inference, logic based languages (KIF), Semantic lexicon

Moreover, the ontology learning implication is evident in the field of text mining [3.32]. [3.42], [3.33], [3.36], [3.37]. The next section presents ontology learning implication in the field of text mining that has become an important research area as a result of the explosion of textual data on the Web, such as emails, blogs, communities, and discussion boards.

3.4. Ontology Learning Process in Text Mining

The traditional Bag-of-Word (BoW) approach that is widely used in the field of text mining provides a simplified mathematical model of the text document without considering the logical connection between the words and sentences [3.3], [3.26], [3.3], [3.32], [3.59]. The BoW approach comprises two major steps described as: (i) implementation of linguistics techniques (Section 2.1.1) for the analysis of the grammatical structure of a sentence and (ii) implementation of statistical techniques that evaluate the occurrence of words in a text document. BoW prunes the infrequent words from the document and chose the words that have high mutual information with the target concept [3.55]. The document representation in the BoW approach is based on the frequency of the occurrence of a word. In the past, various approaches have been proposed to compute occurrence of a word in a document, such as Document Frequency (DF), Inverse Document Frequency (IDF), latent semantic indexing, and Information Gain (IG) [3.52], [3.23], [3.56].

The scientific literature highlights various deficiencies in the BoW approach due to the complex inherent nature of unstructured textual data [3.42], [3.7], [3.59], [3.60]. For instance, BoW fails to understand the underlying meaning of the textual sentence when a single word is ambiguous or out of the context [3.33]. The logic extraction process with the BoW approach becomes more complex when a sentence has multi-word expressions, such as synonyms, polysemy, and homonymy [3.7], [3.36]. The complexity is due to the fact that BOW approach

categorizes textual data without considering the logical connection among words in a sentence [3.62]. Therefore, the results based on BoW do not exceed a 55-60% success rate in terms of accuracy because of the lack of in-depth understanding of textual data [3.34]. However, incorporating ontologies into the traditional BoW approach by utilizing the ontology learning process can improve the text mining process and yield better results [3.32], [3.11], [3.9], [3.51]. Based on the abovementioned facts, a novel approach was suggested that retrieves the semantically relevant information from the multi-sentence sections comprising medical terms [3.42], [3.50]. The approach combined ontology using the Semantic Web Rule Language (SWRL) that contains semantic interpretation of different medical terms. The result of proposed approach indicated that ontology-based approaches have higher precision (76%) than does the traditional BoW term-based approach (68%). Punitha et al. presented a novel comparison of traditional clustering systems with ontology-based document clustering and demonstrated significant improvement in the clustering results [3.37]. Furthermore, Berdnt et al. [3.45] suggested a hybrid approach based on the graph measure computation to calculate the term importance over an entire ontology and further used the measure in the statistical text mining process. The authors used PageRank and HITS algorithms for the term importance measurement and claimed consistent improvement and accuracy as compared with using pure statistical techniques of text mining [3.45]. Similarly, J.Ma et al. proposed a novel Ontology-based Text Mining Method (OTMM) to cluster research proposals based on their similarities in a particular research area [3.32]. The OTMM consisted of a set of concepts, axioms, and relationships that presents an agreed-upon conceptualization of a particular research topic, which automate the process of the categorization of research proposals. The results show significant improvement in the clustering of the similar research proposals. Another interesting study was conducted by

Lula et al. [3.64]. The authors suggested a clustering-based technique that determines similarity between objects based on their relationships and used an ontology schema with agglomerative hierarchical clustering algorithm [3.64]. In the proposed technique, every object is presented as: (i) object description (object name, relationships between objects) and (ii) category description (hierarchal schema, class). The ontology-based clustering technique demonstrated significant improvement. An ontology learning-based text mining architecture is presented in Figure 9 that incorporates ontology learning into the preprocessing phase of text mining, which organizes the text documents into a fix number of predefined categories [3.34], [3.60]. The ontology learning process extracts the concepts from the text and stores it into the ontology repository. Furthermore, the relevant ontologies are mapped with the texts that are used in different machine learning algorithms for the text categorization and provide semantic-based text interpretation to the text document. Table 4 summarizes various successful attempts of different researchers to use ontologies in the field of text mining. We can observe from the Table 4 that most of the existing researchers are emphasizing on implementing the linguistic and statistical techniques in ontology-based text mining approach. However, incorporating the semantic-based techniques is still an open issue for the future work.

3.5. Developments in Ontology Learning Techniques

The ontology learning community is focusing on improving the efficiency of relation discovery techniques and the authenticity of ontology learning evaluation approaches. Moreover, the learning of ontologies from Web-based textual data and different language formats has also been a topic of great interest in the past few years [3.69], [3.70], [3.78], 3.68]. The progress in the aforementioned important fields will be presented in the following subsections.

Table 4. Ontology-based text-mining (case studies)

Ref.	Ontology-Based Text Mining				
	Ontology	Techniques			Source Used
		Linguistic-based	Statistical-based	Semantic-based	
[3.44]	Medical Ontology MeSH	Tokenization, Stemming	Path-base, Feature-based, Information content-based similarity measure, K-mean Clustering	-	PubMed, MEDLINE doc
[3.31]	OntoGen	POS tagging, Parsing, removal of punctuation, stemming	Conceptual clustering, Concept mapping, Information Gain measures, Support Vector Machine, Decision tree, K-nearest neighbor, Nave-Bayes	Prolog	JSTOR E-economic Document
[3.32]	Research Ontology	Tokenization, Syntactic structure dependency analysis	Agglomerative clustering, TFIDF, WordNet, LSI, Self-Organizing Map	-	Research Papers
[3.45]	Ontological Network	Stemming, Tokenization, Parsing,	Agglomerative clustering, PAGERANK, HITS, WordNet, TFIDF, Classification, Clustering	-	Systematized nomenclature of medicine clinical terms (SNOMED CD)
[3.9]	Digital library Ontology	Syntactic structure dependency analysis	TFIDF, Agglomerative clustering	Inductive algorithm, Logic inference system	Digital Library
[3.82]	Agricultural Ontology	Tokenization, Stemming	K-Nearest neighbor, Nave-Bayes, Decision tree	Case-based reasoning (CBR)	Agricultural Knowledge Database
[3.79]	UMLS	GENIA POS tagger,	Multiple Classification Ripple Down Rules (MCRDR)	-	Biomedical natural language documents
[3.80]	OntoPlus	Tokenization, Stop-word removal, Stemming	Co-occurrence analysis	-	Cyc Knowledge base
[3.81]	Domain Ontology	POS, Palavras parser,	Decision tree	Inference system, Description Logic	Natural Language Text Document

3.5.1. Ontology Learning Evaluation

The evaluation process of ontology learning assists users to select the finest ontology that suite the user's requirements. The ontology learning process can be evaluated by utilizing various objective measures, such as consistency, accuracy, and completeness [3.69]. Moreover, the ontology can be evaluated with respect to specific domain or application areas. For instance, in a textual information retrieval system, the main objective of evaluation is to provide a document that satisfies user's queries. Moreover, if the resulting ontology is insufficient for a specific application, then the malfunctioning module of ontology learning process should also be identified by utilizing ontology learning evaluation methods. In the past few years, ontology evaluation methods have been categorized into various groups. The categorization is based mainly on the purpose of evaluation [3.70] and kind of ontology [3.69] being evaluated. Various evaluation methods are presented in the subsequent text.

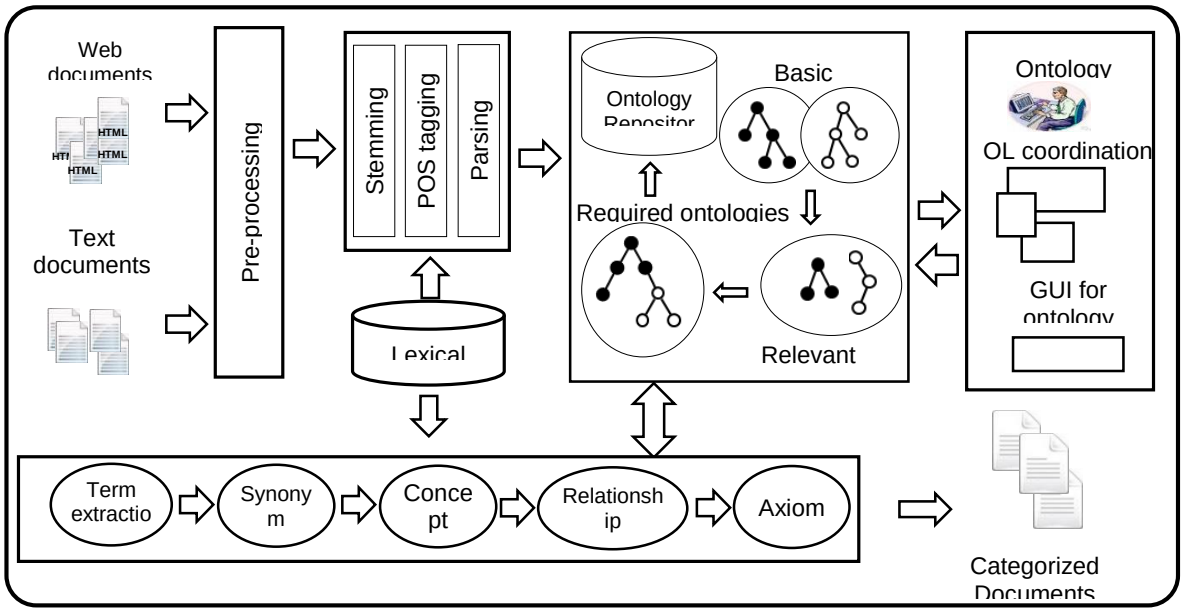

Figure 9. Ontology-based text mining

3.5.1.1. Task-based Approach

The task-based evaluation estimates the accuracy of ontologies by considering a specific application. For instance, in the case of textual document retrieval, task-based evaluation assesses the quantitative and qualitative adequacy of retrieved documents in response to users' queries. Several scientific articles discuss the use of task-based evaluation approach for evaluating the ontologies. For instance Christopher et al. [3.69] utilized task-based evaluation to assess the performance of a Web search engine by comparing example queries with the actual Web search results. Similarly, Clarke et al. [3.70] used the same approach for gene ontology evaluation. Every task-based evaluation is unique and designed for a specific application. Therefore, no finite set of measures and standardized rules can be defined [3.70].

3.5.1.2. Corpus-based Approach

Such approach utilizes domain-specific data source to evaluate the extent to which the resulting ontologies cover the corresponding domain. For instance, in the textual information retrieval process, the ontology is compared with the basic terminologies associated with a text corpus. The text corpus can be considered as a set of facts. The corpus-based approach evaluates the resemblance between the set of facts and the patterns that are logically derived from the ontologies [3.78].

3.5.1.3. Expert-based Approach

Such evaluation is performed by a team of domain experts who evaluate the ontologies based on standards and predefined criteria for a specific domain. However, the adequacy of such approach depends on cost, competency, and availability of domain experts [3.78].

3.5.1.4. Gold Standard Approach

In the gold standard approach, the quality of the ontology is evaluated by comparing the ontology with the manually build predefined "gold standard" ontology. The gold standard ontology is designed by a single expert or a team of domain experts. A typical gold standard evaluation approach comprises three major steps. First, a single expert or a team of domain experts manually build the gold standard ontology. Second, some concepts, rules, and relations are deliberately removed from the gold standard ontology. Third, the modified ontology is augmented with the ontology learning process. The degree to which the ontology learning process manages to reconstruct the modified ontology is the measuring criterion for the gold standard approach [3.70], [3.78].

3.5.1.5. Level-based Approach

Due to complex nature of ontologies, fairly good choice is to focus on the evaluation of different levels (modules) of ontology rather than evaluating the entire ontology. In the scientific literature, different levels of ontologies are described, such as: (i) lexical, (ii) taxonomy, (iii) semantic, (iv) context, (v) syntactic, and (vi) architecture. The lexical level presents the vocabularies (terms) used to identify the concept in the ontology. Evaluation at the lexical level comprises comparison of the terms with the domain-specific text corpora. The taxonomy and semantic levels of ontology show the relationship between the terms that makes a concept. The ontologies usually are described in a formal language, such as Web Ontology Language (OWL) and Resource Description Framework (RDF) [3.77]. The syntactic level of ontology presents the syntactic structure, such as keywords of the language. Unlike the rest of the described levels, architecture level focuses on decided design principles or criteria that are identified prior to the design ontology. An evaluation at the architecture level measures the extent to which the

ontology fulfills the predefined design principles [3.77]. Table 5 presents the relationship between ontology evaluation approaches and the level-based approach.

The scientific literature shows that approach: (a), (b), (c), and (e) discussed in aforementioned text exhibit considerable improvements in the ontology learning evaluation process. However, in the case of large-scale textual data and frequent evaluation process, approach (d) is practically the most feasible method [3.76], [3.77]. Nevertheless, the absence of a well-founded and standard evaluation model in current gold-standard based evaluation has been observed in recent studies [3.76], [3.77].

Table 5. Overview of approaches to ontology evaluation on different levels

Level-based Approach	Ontology Evaluation Approach			
	Task-based Assessment	Corpus-based Assessment	Gold-Standard	Expert-based Assessment
Lexical, data	✓	✓	✓	✓
Semantic relations	✓	✓	✓	✓
Taxonomy, hierarchy	✓	✓	✓	✓
Context (application)	✓	✓		✓
Syntactic	✓			✓
Architecture, design, structure	✓			✓

3.5.2. Ontology Learning from Social Data

Recently, researchers investigated the importance of the ontology learning process for acquiring social data from Web, such as blogs and Wikis. Kotis et al. [3.71] presented two techniques for automatically building the ontologies from social data on the Web. Moreover, the authors evaluated the proposed ontology learning techniques by utilizing Yahoo! and Google query datasets. Similarly, Mika et al. [3.72] proposed an abstract model of semantic social

networks termed as “actor concept instance model” that presents a simple graph transformation by highlighting ontologies of concepts and social networks of users on the Web. Recently, Brisson et al. also augmented corpus-based ontology learning with collaborative tagging systems, social networking platforms, and micro-blogging services [3.73].

The intertwining of ontology learning and large-scale textual data available on the Web is a need of the current era. However, the potential growth of Web data will introduce new challenges in ontology learning. For instance: (i) the fact that Web data is prone to spelling errors and grammatical errors in the text, possibly leading to build incorrect ontologies, (ii) the issue of authority and validity of contents available on the Web for ontology building process, and (iii) the representation of ontologies as a language-independent construct on the Web.

3.6. Current Issues and Future Directions and for Ontology Learning in Text Mining

Based on the preceding discussion, we can conclude that ontologies are becoming an important factor for semantic information annotation from textual data. However, the ontology learning process still faces issues and challenges that need to be addressed. These issues are outlined below and in our opinion the most important directions for future work in the field of ontology learning.

a. Fully automated ontology learning systems: Construction of fully-automated ontologies from text is still not feasible. Ontology represents the conceptualization of a specific domain. Human validation and consensus are always required for the construction of ontology. An automated ontology learning system would eliminate the need for human intervention during the construction process, which currently creates significant difficulties in reaching a consensus regarding the concepts, definitions, and relations used [3.48], [3.54]. This direction

requires more research in order to provide viable solutions for an automated ontology learning system.

b. Ontology learning evaluation: Evaluation is an important issue in the ontology learning process. Various ontology evaluation techniques have been discussed in Section 4.1. However, the majority of such evaluation techniques are either domain specific or application specific. Evaluation of ontology is critical because of the absence of standard rules that can apply to all ontologies associated with the textual data. Moreover, designing a formal method for evaluating the ontology learning process is still an open problem [3.12], [3.39], [3.48], [3.66]. Furthermore, standard benchmarks are required for evaluating ontologies from dynamic domain areas [3.12].

c. Portability of domain ontologies across different platforms: Researchers have focused on building domain ontologies that rely solely on the language-dependent domain-specific environment. Integration of information from different languages can improve the outcomes of the ontology learning process. However, portability across multilingual platforms and mapping of such ontologies into different domain ontologies are still open issues for future research. Moreover, incorporating new changes in existing ontologies and mapping new ontologies with the existing ones are still challenging issues for researchers.

Ontology learning in text mining can clearly benefit from the automation of the ontology building process from text by utilizing formal evaluation methods and mapping portable ontologies via multilingual platforms. Moreover, a closer look into previous research related to the ontology learning process revealed some additional consensus that requires more attention. The main observations are that: (i) because of the availability of multiple conflicting or complementing ontologies, standard methods are required to determine the correspondence

between the relations, terms, and concepts in different ontologies, (ii) scalability and robustness issues in Web-scale ontology learning process should also be emphasized, and (iii) ontology maintenance is an important open issue. The ontologies related to a specific domain area dynamically evolve over time. Therefore, the ontologies need to be updated and maintained periodically. Reusing and reformulating the existing ontology instead of building an entire ontology from scratch can be an attractive research project for future researchers. All the above mentioned research directions raise many challenging issues which are still open problems to be solved in the future.

3.7. Conclusions

Despite extensive research in the field of text mining, semantic-based interpretation of unstructured text remains a challenging problem because of the complex nature of textual data. Ontologies can represent efficient textual realization of a domain by utilizing background hierarchical knowledge that is based on meaningful associations between the concepts and objects in the textual documents. Moreover, the ontology-based text interpretation can improve the text mining approach and can yield better results compared with traditional text mining approaches, such as BoW. However, ontology building is a challenging task that currently requires human intervention and iterative processes for design, implementation, and evaluation. Therefore, semi-automatic approaches are becoming available that may ease this burden. This survey presents a comprehensive review of linguistic, statistical, and semantic-based ontology learning techniques. Moreover, the survey describes the implication of ontology learning techniques on different ontology learning process modules, such as term, concept, and axiom. Furthermore, the discussion presents the ontology-based text mining architecture as an example of ontology learning implication into the field of text mining, the most promising research area

for logical interpretation of the text corpora. The survey also reviews the major issues and challenges in the ontology learning process, such as fully automated ontology learning systems, ontology learning evaluation, and portability of domain ontologies across different platforms.

3.8. Acknowledgments

This material was based upon work supported by the U.S. Department of Energy, Office of Science, under Contract DE-AC02-06CH11357.

3.9. References

[3.1] Hsieh, S., Lin, H., Chi, N., Chou, K., and Lin, K.: 'Enabling the Development of Base Domain Ontology through Extraction of Knowledge from Engineering Domain Handbooks', Advanced Engineering Informatics, 2011, 2(25), pp. 288-296

[3.2] Marinica, C., Guillet, F.: 'Knowledge-Based Interactive Post-mining of Association Rules using Ontologies', Knowledge and Data Engineering, 2010, 6(22), pp.784-797

[3.3] Khalida, B., Adil. T.: 'Lightweight Domain Ontology Learning from Texts: Graph Theory-based Approach using Wikipedia', International Journal of Metadata, Semantics and Ontologies, 2014, 9(2), pp. 83-90

[3.4] Fanghuai, H., Zhiqing, S., Tong, R.: 'Self-Supervised Chinese Ontology Learning from Online Encyclopedias', The Scientific world Journal, 2014, pp. 34-48

[3.5] Fernandez, J.D., Gutierrez, C., and Martinez, M.A.: 'RDF Compression: Basic Approach'. Proc. Int. Conf. World Wide Web, 2010, pp. 1091-1092

[3.6] Wong, W., Liu, W., and Bennamoun, M.: 'Ontology Learning from Text: A Look Back and into the Future', ACM Journal Computing Survey, 2012, 44(4), pp. 12-32

[3.7] Poelmans, J., Ignatov, D., Viaene, S., Dedene, G., and Kuznetsov, S.: 'Text Mining Scientific Papers: A Survey on FCA-based Information Retrieval Research', Language Resources and Evaluation, 2012, 46(2), pp. 155-176

[3.8] He, J., Ren, K., and Yu, W.: 'An Improved Supervised Word Sense Disambiguation Method in the Biomedical Domain', Applied Mechanics and Materials, 2013, 241(244), pp. 3103-3106

[3.9] Madhu, G., Govardhan, A., and Rajinikanth, T.V.: 'Intelligent Semantic Web Search Engines: A Brief Survey', International Journal of Web & Semantic Technology, 2011, 2(1), pp. 34-42

[3.10] Ivo, S., Rosario, G., Paulo, N.: 'Evaluating Techniques for Learning Non-Taxonomic Relationships of Ontologies from Text', Expert System with Applications, 2014, 41(11), pp. 5201-5211

[3.11] Kulick, S., Bies, A., and Maamouri, M.: 'Consistent and Flexible Integration of Morphological Annotation in the Arabic Treebank'. Proc. Int. Conf. Language Resources and Evaluation (LREC), 2010, pp. 1- 8

[3.12] Richard, G., Maria, J.: 'SMOL: A Systemic Methodology for Ontology Learning from Heterogeneous Sources', Journal of Intelligent Information Systems, 2014, 42(3), pp. 415-455

[3.13] Scanniello, G., Risi, M., and Tortora, G.: 'Architecture Recovery using Latent Semantic Indexing and K- Means: An Empirical Evaluation'. Proc. Int. Conf. Software Engineering and Formal Methods (SEFM), Canada, 2010, pp. 103-112

[3.14] Sidiahmed, K., Toumouh, A., and Maliki, M.: ‘Effective Ontology Learning: Concept Hierarchy Building using Plain Text Wikipedia’. Proc. Int. Conf. Web and Information Technology (ICWIT), Algeria, 2012, pp. 171- 178

[3.15] Hu, K., Tian, Y., and Wang, Y.: ‘Semantic Manipulations and Formal Ontology for Machine Learning based on Concept Algebra’, International Journal of Cognitive Informatics and Natural Intelligence, 2011, 3(5), pp. 46-67

[3.16] Bateman, J., Hois, J., Ross, R., and Tenbrink, T.: ‘A Linguistic Ontology of Space for Natural Language Processing’, Artificial Intelligence, 2010, 14(174), pp.1027-1071

[3.17] Hongsheng, X., Ruiling, Z.: ‘Research on Data Integration of the Semantic Web based on Ontology Learning Technology’, TELKOMNIKA Indonesian Journal of Electrical Engineering, 2014, 12 (1), pp. 167-178

[3.18] E. Drymonas, E., Zervanou, K., and Petrakis, E.G.M.: ‘Unsupervised Ontology Acquisition from Plain Text: the OntoGain System’. Proc. Int. Conf. Natural Language Processing and Information System, 2010, pp. 277-287

[3.19] Cunninghum, H., Tablan, V., Roberts, A., and Bontcheva, K.: ‘Getting More Out of Biomedical Documents with GATE’s Full Lifecycle Open Source Text Analytics’, PLoS Computational Biology, 2013, 9(2), pp. 234- 255

[3.20] Zaod, N and Lau, S.: ‘Emerging of Academic Information Search System with Ontology-based Approach’, Procedia-Social and Behavioral Sciences, 2014, 116, pp. 132-138

[3.21] Antoy, S., and Hanus, M.: ‘Functional Logic Programming’, Communications, 2010, 53(4), pp. 74-85

[3.22] Panchenko, A., and Morozova, O.: ‘A Study of Hybrid Similarity Measures for Semantic Relation Extraction’, Proc. Int. Conf. Innovative Hybrid Approach to the Processing of Textual Data, 2010, pp. 10-18

[3.23] Turney, P., and Pantel, P.: ‘From Frequency to Meaning: Vector Space Models of Semantics’, Journal of Artificial Intelligence, 2014, 37, pp. 141-188

[3.24] Kavitha, V., and Punithavall, M.: ‘Clustering Time Series Data Stream- A Literature Survey’, International Journal of Computer Science and Information Security, 2010, 8(1), pp. 289-294

[3.25] Sathiyakumari, K. and Manimekalai, G.: ‘A Survey on Various Approaches in Document Clustering’, International Journal of Computer Technology and Application (IJCTA), 2011, 2(5), pp. 1534-1539

[3.26] Yonghong, Y., and Bai, W.: ‘Text Clustering based on Term Weights Automatic Partition’. Proc. Int. Conf. Computer and Automation Engineering (ICCAE), China, 2010, pp. 373-377

[3.27] Wimalasuriya, D.C, and Dou, D.: ‘Ontology-based Information Extraction: An Introduction and a Survey of Current Approaches’, Journal of Information Science, 2010, 36(3), pp. 306-323

[3.28] David, D.: ‘Adaptive Ontologies through Social Evolution’, Proc. Int. Conf. on Autonomous Agents and Multi-Agent Systems, 2014, pp. 1743-1744

[3.29] Pahlavi, N., and Muggleton, S.: ‘Towards Efficient Higher-Order Logic Learning in a First-Order Datalog Framework, 2012, Latest Advances in Induction Logical Programming, Imperial College Press

[3.30] Suyanto, and Pancaputra, M.A., and Wahyuono, R.A.: 'Designing of Expert System for Troubleshooting Diagnosis on Gas Chromatagraphy GC-2010 by Means of Inference Methods', Proc. Int. Conf. on Uncertainty Reasoning and Knowledge engineering (URKE), 2014, 1, pp. 5-8

[3.31] Sergeja, V., and Zoran, B.: 'Ontology-based Multi-Label Classification of Economic Article', Computer Science and Information Systems (ComSIS), 2011, 1(1), pp. 101-119

[3.32] Ma, J., Xu, W., Sun, Y., Turban, E., Wang, S., and Liu, O.: 'An Ontology-based Text Mining Method to Cluster Proposals for Research Project Selection', IEEE Transaction on System, Man, and Cybernetics, 42(3), 2012, pp. 784-790

[3.33] Wu, L.,Hoi, S. C., and Yu, N.: 'Semantics-Preserving Bag-of-Words Models and Applications', Image Processing, 2010, 7(19), pp. 1908-1920

[3.34] Aggarwal, C., and Zhai, C.: 'A Survey of Text Classification Algorithms', pp. 77-121 in Aggarwal, C., and Zhai, C. (Eds): 'Mining Text Data' (Springer 2010).

[3.35] Tanmay, B., and Murthy, C. A.: 'CUES: A New Hierarchical Approach for Document Clustering', Journal of Pattern Recognition Research, 2014, 8(1)

[3.36] Jiang, X., and Tan, A.: 'CRCTOL: A Semantic-based Domain Ontology Learning System', Journal of the American Society for Information Science and Technology, 2010, 1(61), pp. 150-168

[3.37] Punitha, S.C., Mungunthadevi, K., and Punithavalli, M.: 'Impact of Ontology based Approach on Document Clustering', International Journal of Computer Applications, 2011, 22(2), pp. 22-26

[3.38] Subhashini, R., and Akilandeswari, J.: ‘A Survey on Ontology Construction Methodologies’, International Journal of Enterprise Computing and Business Systems, 2011, 1(1)

[3.39] Zavitsanos, E., Paliouras, G., and Vouro, A.: ‘Gold Standard Evaluation of Ontology Learning Methods through Ontology Transformation and Alignment’, Knowledge and Data Engineering, 2011, 11(23), pp. 1635-1648

[3.40] Fanghuai, H., Zhiqing, S., and Tong, R.: ‘Self-Supervised Chines Ontology Learning from Online encyclopedias’, The Scientific World Journal, 2014, 13(3), pp. 234-250

[3.41] Serra, I., Rosario, G., and Paulo, N.: ‘PARNT: A Statistic based Approach to Extract Non-Taxonomic Relationships of Ontologies from Text’, Proc. Int. Conf. Information Technology: New Generations (ITNG), Las Vegas, 2014, pp. 561-566

[3.42] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm, Accessed September 2014

[3.43] http://pypi.python.org/pypi/nltk, Accessed September 2014

[3.44] Zhang, X., Jing, J., Hu, X., Ng, M., and Zhou, X.: ‘A Comparative Study of Ontology based Term Similarity on PubMed Document Clustering’. Proc. Int. Conf. Database Systems for Advanced Applications, pp. 115-126

[3.45] Berndt, D. J., McCart, A., and Luther, S.: ‘Using Ontology Network Structure in Text Mining’, AMIA Annu Symp Proc. 2010, pp.41-45

[3.46] Tang, S. L., Ip, W., and Tsang, H. C.: ‘Is Naïve Bayes a Good Classifier for Document Classification?’, Journal of Software Engineering and Its Applications, 2011, 5(3), pp. 37-46

[3.47] Malhotra, A., Younesi, E., Gundel, M., Muller, B., Heneka, and M., Apitius, M. H.: 'ADO: A Disease Ontology Representing the domain Knowledge Specific to Alzheimer's Disease', Alzheimer's Dementia, 2013, 6, pp. 1-9

[3.48] Shvaiko, P., and Trento, J.: 'Ontology Matching: State of the Art and Future Challenges', IEEE Transactions on Knowledge and Data Engineering, 2011, 25(1), pp. 158-176

[3.49] Glimm, B., Horrocks, I., Motik, B., Shearer, R., and Stoilos, G.: 'A Novel Approach to Ontology Classification', Web Semantics: Science, Services and Agents on the World Wide Web, 2012, 14, pp. 84-101

[3.50] Nan, L., and Desheng, W. D.: 'Using Text Mining and Sentiment Analysis for Online Forums Hotspots Detection and Forecast', Decision Support Systems, 2010, 48(2), pp. 354-368

[3.51] Jian, M., Wei, X., Sun, Y. H., Turban, E., Wang, and S., Liu, O.: 'An Ontology-based Text-Mining Method to Cluster Proposals for Research Project Selection', IEEE Transactions on Systems, Man and Cybernetics, Part A: System and Humans, 2012, 42(3), pp. 784-790

[3.52] Krishnaiah, V., Narsimha, G., and Chandra, N. S.: 'Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques', International Journal of Computer Science and Information Technologies (IJCSIT), 2014, 4(1), pp. 39-45

[3.53] Sharma, R., Malik, B., and Ram, A.: 'Local Density Differ Spatial Clustering in Data Mining', International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), 2013, 3(3), pp. 393-397

[3.54] Asuncion, G. P., Marcos, M. R., Alejandro, R. G., Vazquez, V. G., and Jose, M.: 'Ontologies in Medicinal Chemistry: Current Status and Future Challenges', Current Topics in Medicinal Chemistry, 2013, 13(5), pp. 576-590

[3.55] Xiong, H., Sun, S., and Feng, Y.: 'Text Clustering based on Kernel KNN Clustering Algorithm', International Journal of Applied Mathematics and Statistics, 2013, 46(16)

[3.56] Zhang, H., Ho, J. K. L., Wu, Q. M. J., and Ye Y.: 'Multidimensional Latent Semantic Analysis Using Term Spatial Information', IEEE Transaction on Cybernetics, 2013, 43(6), pp. 1625-1640

[3.57] Patil, L. H., and Atique, M.: 'A Novel Feature Selection based on Information Gain Using WordNet', Proc. Int. Conf. Science and Information Conference (SAI), 2013, pp. 625-629

[3.58] Jiang, J. and Ling,T.: 'Mahalanobis-Taguchi System and Selective Naïve Bayesian Algorithm for Multivariate Pattern Recognition', Advanced Science Letters, 2013, 19(2), pp. 638-641

[3.59] Riadh, B., Abir, M., and Jalel, A.: 'Using a Bag of Words for Automatic Medical Image Annotation with a Latent Semantic', International Journal of Artificial Intelligence and Applications (IJAIA), 2013, 4(3)

[3.60] Susana, A., and Maria, V.: 'Texton Theory Revisited: A Bag-of-Words Approach to Combine Textons', Pattern Recognition, 2012, 45(12), pp. 4312-4325

[3.61] Georgiev, G., Zhikov, V., Osenova, P., Simov, K., and Nakov, P.: 'Feature-Rich Part-of-Speech Tagging for Morphologically Complex Language: Application to Bulgarian', Proc. Int. Conf. Association for Computational Linguistics, pp. 492-502

[3.62] Ray, S., and Chandra, N.: ‘Domain based Ontology and Automated Text Categorization based on Improved Term Frequency-Inverse Document Frequency’, International Journal of Modern Education and Computer Science (IJMECS), 2012, 4(4), pp. 28-35

[3.63] Gil, R. J., M. Bautista, M. J.: ‘A Novel Integrated Knowledge Support System based on Ontology Learning: Model Specification and a Case Study’, Knowledge-Based System, 2012, 36, pp. 340-352

[3.64] Lula, P., and Pekosz, G. P.: ‘An Ontology-based Clustering Analysis Framework’, Pro. Int. Conf. on Ontology Supported Business Intelligence, pp. 23-46

[3.65] Hung, C., and Lin, H.K.: ‘Using Objective Words in SentiWordNet to Improve Word-of-Mouth Sentiment Classification’, IEEE Intelligent Systems, 2013, 28(2), pp. 47-54

[3.66] Murdock, J., Buckner, C., and Allen, C.: ‘Evaluating Dynamic Ontologies’, Communication in Computer and Information Science, 2013, pp. 258-275

[3.67] Fermin, L. Cruz, J. A. Troyano, B. pontes, F. J. Ortega.: ‘Building Layered Multilingual Sentiment Lexicons at Synset and Lmma Levels’, Expert System with Applications, 2014, 41(13), pp. 5984-5994.

[3.68] Navigli, R., and Ponzetto, S. P.: ‘BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network’, Artificial Intelligence, 2012, 193, pp. 217-250

[3.69] Welty, C., Kalra, R., and Chu-Carrol, J.: ‘Evaluating Ontological Analysis’, Proc. ISWC-03 Workshop on Semantic Integration, 2003.

[3.70] Clarke, E. L.: ‘A Task-based Approach for Gene Ontology Evaluation’, Journal of Biomedical Semantics, 4, 2013

[3.71] Kotis, K., and Pasalouros, A.: ‘Automated Learning of Social Ontologies’ in ‘Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances’, (W. Wong, W. Liu, and M. Bennamoun, edn.), IGI Global, Hershey, 2011

[3.72] Mika, P.: ‘Ontologies Are Us: A Unified Model of Social Networks and Semantics’, Web Semantic, 5(1), 2007, pp.5-15

[3.73] Brisson, L., and Collard, M.: ‘An Ontology Driven Data Mining Process’, Proc. Int. Conf. Enterprise Information Systems, 2013, pp. 213-225

[3.74] Agrawal, R., Imieliński, T., and Swami, A.: ‘Mining Association Rules between Sets of Items in Large Databases”, Proc. Int. Conf. ACM SIGMOD, Management of Data, 2012

[3.75] Antinio, L., Teflioudi, C., Hose, K.: ‘AMIE: Association Rule Mining under Incomplete Evidence in Ontological Knowledge Base’, Proc. Int. Conf. World Wide Web, 2013, pp. 413-422

[3.76] Zavitsanos, E., Paliouras, G., and Vouros, G.A.: ‘Gold Standard Evaluation of Ontology Learning Methods through Ontology Transformation and Alignment’, IEEE Transactions on Knowledge and Data Engineering, 2011, 23(11), pp.234-248

[3.77] Dellschaft, K., Staab, S.: ‘On How to Perform a Gold Standard based Evaluation of Ontology Learning’, Proc. Int. Conf. on the Semantic Web, 2010, pp. 228-241

[3.78] Petasis, G., Karkaletsis, V., and Paliouras G.: ‘Ontology Population and Enrichment: State of the Art’, Knowledge-Driven Multimedia Information Extraction and Ontology Evaluation, 2014, pp. 134-166

[3.79] Juana, M., Rafael, V., Jesualdo, F., Francisco, S., and Rodrigo, B.: ‘Ontology Learning from Biomedical Natural Language Documents Using UMLS’, Expert Systems with Application, 2011, 38, pp. 12365-12378

[3.80] Inna, N., Dunja, M., and Luke, B.: ‘OntoPlus: Text-driven Ontology Extension Using Ontology Content, structure, and Co-occurrence Information’, Knowledge-based Systems, 2011, 24, pp. 1261-1276

[3.81] Demnis, A., Sandro, R., Carolina, M., and Rove, C.: ‘Automatic Information Extraction from Texts with Inference and Linguistics Knowledge Acquisition Rules’, Proc. Int. Conf. Web Intelligence and Intelligent Agent Technology, 2014, pp. 151-154

[3.82] Zheng, lu., He, yun, Qian, L.: ‘Construction of the Ontology-based Agricultural Knowledge Management System’, Journal of Integrative Agriculture, 2012, pp. 700-723

[3.83] Karkar, A., Saleh, M., Saad, S., and Aljaam, J.: ‘An Arabic Ontology-based Learning System for Children with Intellectual Challenges’, Proc. Int. Conf. Global engineering Education, 2014, pp. 670-675

4. MOBICONTEXT: A CONTEXT-AWARE CLOUD-BASED RECOMMENDATIONS FRAMEWORK

4.1. Abstract

In recent years, recommendation systems have seen significant evolution in the field of knowledge engineering. Most of the existing recommendation systems based their models on collaborative filtering approaches that make them simple to implement. However, performance of most of the existing collaborative filtering-based recommendation system suffers due to the challenges, such as: (a) cold start, (b) data sparseness, and (c) scalability. Moreover, recommendation problem is often characterized by the presence of many conflicting objectives or decision variables, such as users' preferences and venue closeness. In this paper, we proposed MobiContext, a cloud-based Bi-Objective Recommendation Framework (BORF) for mobile social networks. The MobiContext utilizes multi-objective optimization techniques to generate personalized recommendations. To address the issues pertaining to cold start and data sparseness, the BORF performs data preprocessing by using the Hub-Average (HA) inference model. Moreover, the Weighted Sum Approach (WSA) is implemented for scalar optimization and an evolutionary algorithm (NSGA-II) is applied for vector optimization to provide optimal suggestions to the users about a venue. The results of comprehensive experiments on a large-scale real dataset confirm the accuracy of the proposed recommendation framework.

4.2. Introduction

The ongoing rapid expansion of the Internet and easy availability of numerous e-commerce and social networks services, such as Amazon, Foursquare, and Gowalla, have resulted in the sheer volume of data collected by the service providers on daily basis. The continuous accumulation of massive volumes of data has shifted the focus of research

community from the basic information retrieval problem to the filtering of pertinent information [1], thereby making it more relevant and personalized to user's query. Therefore, most research is now directed towards the designing of more intelligent and autonomous information retrieval systems, known as Recommendation Systems.

4.2.1. Research Motivation

Recommendation systems are increasingly emerging as an integral component of e-business applications [4.1]. For instance, the integrated recommendation system of Amazon provides customers with personalized recommendations for various items of interest. Recommendation systems utilize various knowledge discovery techniques on a user's historical data and current context to recommend products and services that best match the user's preferences.

In recent years, emergence of numerous mobile social networking services, such as, Facebook and Google Latitude has significantly gained the attraction of a large number of subscribers [4.2], [4.6]. A mobile social networking service allows a user to perform a "check-in" that is a small feedback about the place visited by the user [4.2], [4.3], [4.22]. Large number of check-ins on daily bases results in the accumulation of massive volumes of data. Based on the data stored by such services, several Venue-based Recommendation Systems (VRS) were developed [4.1]–[4.4]. Such systems are designed to perform recommendation of venues to users that most closely match with users' preferences. Despite having very promising features, the VRS suffer with numerous limitations and challenges. A major research challenge for such systems is to process data at the real-time and extract preferred venues from a massively huge and diverse dataset of users' historical check-ins [4.1], [4.3], [4.4]. Further complexity to the problem is added by also taking into the account the real-time contextual information, such as:

(a) venue selection based on user's personal preferences and (b) venue closeness based on geographic information.

4.2.2. Research Problem

In scientific literature, several works, such as [4.1]–[4.6], and [4.10] have applied Collaborative Filtering (CF) to the recommendation problem in VRS. The CF-based approaches in VRS tend to generate recommendations based on the similarity in actions and routines of users [4.3], [4.2], [4.5]. However, despite being less complicated, most CF-based recommendation techniques suffer from several limitations that make them less ideal choice in many real-life practical applications [4.2]. The following are the most common factors that affect the performance of many existing CF-based recommendation systems:

a. Cold start. The cold start problem occurs when a recommendation system has to suggest venues to the user that is newer to the system [4.3]. Insufficient check-ins for the new user results in zero similarity value that degrades the performance of the recommendation system [4.2]. The only way for the system to provide recommendation in such scenario is to wait for sufficient check-ins by the user at different venues.

b. Data sparseness. Many existing recommendation systems suffer from data sparseness problem that occurs when users have visited only a limited number of venues [4.4]. This results into a sparsely filled user-to-venue check-in matrix. The sparseness of such matrix creates difficulty in finding sufficient reliable similar users to generate good quality recommendation.

c. Scalability. Majority of traditional recommendation systems suffer from scalability issues. The fast and dynamic expansion of number of users causes recommender system to parse millions of check-in records to find the set of similar users. Some of the recommendation

systems [4.3], [4.4], [4.24] employ data mining and machine learning techniques to reduce the dataset size. However, there is an inherent tradeoff between reduced dataset size and recommendation quality [4.1].

The immediate effect of the abovementioned issues is the degradation in performance of most of the CF-based recommendation systems. Therefore, it is not adequate to rely solely on simplistic but memory-intensive CF approach to generate recommendations.

4.2.3. Methods and Contributions

In this paper, we propose MobiContext, a cloud-based Bi-Objective Recommendation Framework (BORF) that overcomes the limitations exhibited by CF-based approaches. To address the cold start issues, our framework utilizes the Hub-Average (HA) inference model [4.16] that maintains a pre-computed list of most popular venues in a user's current vicinity. To address data sparseness caused by zero values of similarities, we utilize a metric known as confidence measure that is combined with similarity computations to improve the recommendation quality. The confidence measure defines the conditional probability that two users will show interest in the same set of venues. More precisely, the confidence measure can be expressed as the ratio of the number of venues visited by both users together to the number of venues visited by any one of the two [4.6].

To improve scalability performance, the proposed cloud-based framework follows Software as a Service (SaaS) approach by utilizing a modular service architecture. The primary advantage of this approach is that the proposed framework can scale on demand as additional virtual machines are created and deployed.

We adopt a bi-objective optimization approach that considers the two primary objectives: (a) venue preference and (b) location closeness. Venue preference determines how much the

venue meets the criteria of user's interests, whereas venue closeness indicates how closely a desired venue is located relative to a user's location. The proposed cloud-based MobiContext BORF generates optimized recommendations by simultaneously considering the trade-offs between the aforementioned objectives. In summary, the contributions of our work are as follows.

a. We propose a cloud-based framework consisting of bi-objective optimization methods named as CF-BORF and greedy-BORF. The GA-BORF is also proposed that utilize evolutionary algorithm (NSGA-II) to suggest optimal venues recommendations.

b. To address the issues pertaining to data sparseness and cold start, we formulate confidence measure with similarity computation. We introduce a pre-processing phase that performs data refinement using HA.

c. We perform extensive experiments on our internal OpenNebula cloud setup running on 96 core Supermicro SuperServer SYS-7047GR-TRF systems. The experiments were conducted on real-world "Gowalla" dataset [4.4].

To the best of our knowledge this is the first work to incorporate the bi-objective optimization techniques into VRS. The rest of the paper is organized as follows. Section 2 presents the system overview. In Section 3, we discuss the proposed BORF framework. Section 4 presents the complexity analysis of the proposed framework. In Section 5, we present the performance evaluation with simulation results. The related work is reviewed in Section 6, and Section 7 concludes the paper.

4.3. System Overview

Most of the existing recommendation systems utilize centralized architectures [4.3], [4.4], [4.5], [4.7] that are not scalable enough to process large volume of geographically distributed

data. The centralize architecture for venue recommendations must consider users' preferences, check-in history, and social context simultaneously to generate optimal venue recommendations. Therefore, to address the scalability issue, in the subsequent text we introduce the decentralized cloud-based MobContext BORF approach.

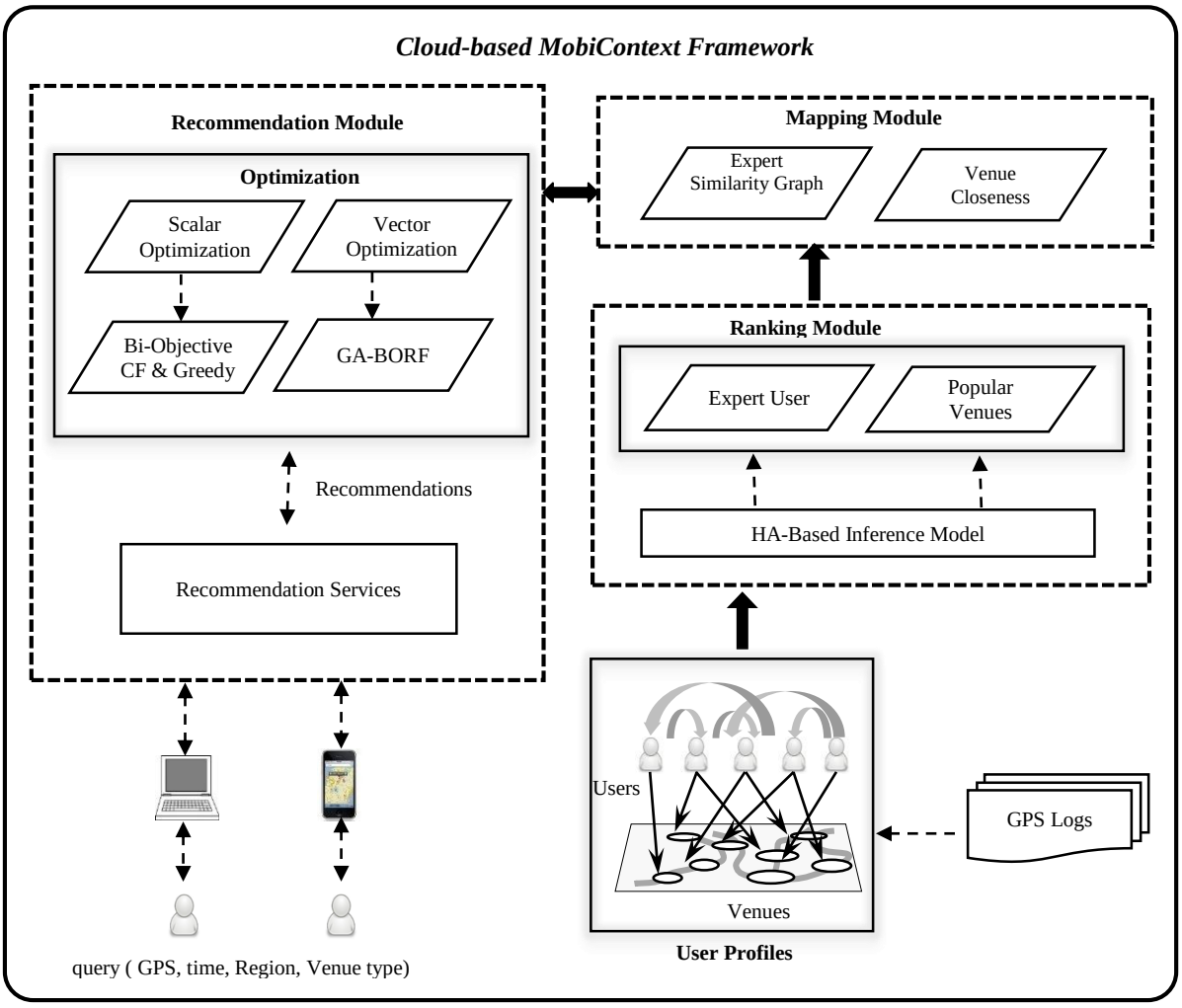

Figure 10. Top level architecture of the cloud-based MobiContext BORF framework

4.3.1. Major Components

As reflected in Figure 10, the MobiContext framework maintains records of user's profiles for each geographical region distributed on the basis of cities. The arrows from users to

venue at lower right of Figure 10 indicate the number of check-ins performed by each user at various venues. A user's profiled consists of the user's identification, venues location visited by the user, types of venues, and time of check-in at a venue. On top of user's profiles, the ranking module performs functionality during the pre-processing phase of data refinement. The ranking module applies HA inference model on users' profiles to assign ranking to the set of users and venues based on mutual reinforcement relationships [4.16]. The idea is to extract a set of popular venues and expert users. We call a venue as popular, if it is visited by many expert users, and a user is expert if (s)he has visited many popular venues [4.15], [4.16]. The users and venues that have very low scores are pruned from the dataset during pre-processing phase to reduce the online computation time.

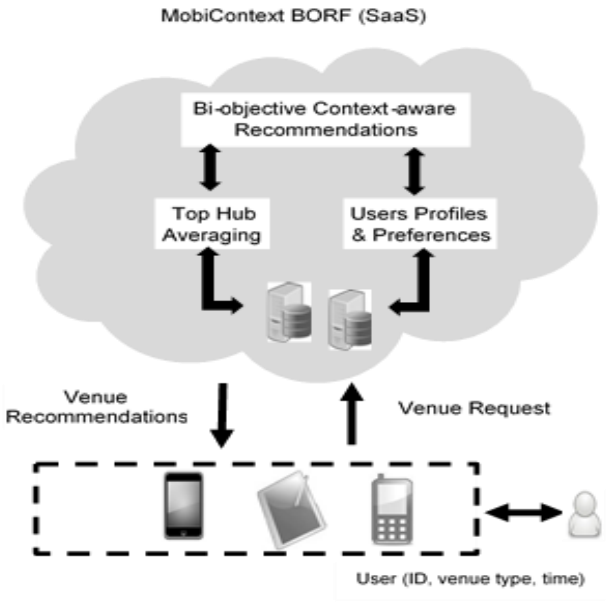

Figure 11. MobiContext cloud-based services mapping

The mapping module computes similarity graphs among expert users for a given region during pre-processing phase. The purpose of similarity graph computation is to generate a network of like-minded people who share the similar preferences for various venues they visit in

a geographical region. Such a network can be useful for generating collaborative recommendation for the other users. The mapping module also computes venue closeness based on geographical distance between the current user and popular venues. The purpose of calculating venue closeness is to suggest venues that are located in the closer vicinity of the user's current location.

At left side of Figure 10 is the recommendation module that runs a service to receive recommendation queries from users. A user's request consists of: (a) current context (such as, GPS location of user, time, and region), (b) venue type (e.g., restaurant, shopping mall, and gas station), and (c) a bounded region surrounding the user from where the top N venues will be selected for the current user. Where N is the user defined parameter [4.7], [4.4], [4.3]. The recommendation service passes the user's query to optimization module that utilizes scalar and vector optimization techniques to generate an optimal set of venues [4.13], [4.14], [4.20]. The detailed description of the aforementioned optimization techniques are presented in Section 3.2. The scalar optimization technique utilizes the CF-based approach and greedy heuristics to generate user preferred recommendations. The vector optimization technique, namely GA-BORF, employs evolutionary algorithms, such as Non-dominated Sorting Genetic algorithm (NSGA-II) to produce recommendations. The venues at the top of the recommended list will be the ones that most satisfy the users' preferences.

4.3.2. MobiContext Cloud-based Services

As reflected in Figure 11, the proposed framework utilizes the SaaS layer of cloud stack to provide real-time personalized recommendations to a user. Users interact with the recommendation framework using mobile devices without requiring significant knowledge about the underlying details of the cloud architecture. The proposed cloud-based MobiContext

framework defines the Service Level Agreement (SLA) as a user's satisfaction level with a recommended venue. To maintain the desirable level of SLA, the framework is designed to recommend only those venues to a given user that are most closely matching with the user's taste as well as are at shortest distance from the user. Moreover, the cloud-based architecture allows the MobiContext framework to scale on runtime. Additional virtual machine can be easily deployed to handle high demand from the user.

Table 6. Notations and their meanings

Symbols	Meaning
$r_{i,v}.r_{j,v}$	Number of check-in at venue *v* performed by the user *i* and *j*
m_c	Venue check-in matrix *m*
V	Set of all venues
E	Set of expert user in a region
∂	Total number of popular venues checked-in by expert users
S	Set of venues visited by expert user *e* but not visited by the current user *c*
p_v	Popular venues
s_e	Similarity value of an expert user e and current user c.
v_c	Venues closeness
$\overline{r_u}$	Average number of check-ins of an active user *u*
$s_r(i,j)$	Similarity matrix of user *i* and *j*
$s_c(u_i, v_j)$	Proximity matrix of user $\boldsymbol{u_i}$ and venue $\boldsymbol{v_j}$
$f_1(i_c)$	Fitness function 1 of individual $\boldsymbol{i_c}$
$f_2(i_c)$	Fitness function 2 of individual $\boldsymbol{i_c}$

4.4. Mobicontext Recommendation Framework

In this section, we discuss in detail the functionality of the proposed MobiContext framework. The most frequently used acronyms in this paper are listed in Table 6. In terms of functionality, MobiContext framework has two main phases: (a) a pre-processing phase and (b) a

recommendation phase. The detailed description of the above mentioned phases is presented in the following subsequent sections.

4.4.1. Pre-processing Phase

The pre-processing is performed offline and is further divided into two phases: (a) ranking phase and (b) mapping phase, as described in the following subsections.

4.4.1.1. Ranking Module

The HA inference model is applied on users' profiles to compute ranking for users and venues. The higher ranked venues and users are known as popular venues and expert users, respectively. To compute the expert users' and popular venues' scores for a region R, the framework will generates region-wise user-to-venue check-in matrix denoted by 〚 M〛_c. Let [p_v] and 〚[e〛_u] represent score matrices for a popular venue and an expert user, respectively, for a region R. The following formulas compute the score for popular venues and expert users:

$$p_v = M_c^T \times e_u. \tag{4.1}$$

$$e_u = M_c \times p_v \times \frac{1}{\partial}. \tag{4.2}$$

If we use $\boldsymbol{p_v^{<n>}}$ and $\boldsymbol{e_u^{<n>}}$ to represent the scores of popular venues and expert users at nth iteration, then the following equations generate the score of popular venues and expert users iteratively.

$$p_v^{<n>} = (M_c^T \times M_c) \times p_v^{<n-1>}. \tag{4.3}$$

$$e_u^{<n>} = (M_c \times M_c^T) \times e_u^{<n-1>}. \tag{4.4}$$

The purpose of using HA method is to generate a subset of users, who have visited popular venues, and a subset of venues that are frequently visited by expert users.

4.4.1.2. Mapping Module

The mapping phase computes the similarity among the expert users (that were generated by the ranking phase) using Pearson Correlation Coefficient (PCC). The PCC is widely used in recommendation systems to generate similarity graphs among users [4.4], [4.15], [4.16], [4.17], [4.19]. The graph constructed in the mapping phase will be made available for online recommendations.

The value of the PCC ranges between -1 and +1, where the value close to 1 indicates the higher degree of similarity exists between two users. If the value of PCC is zero or less than zero, then this means the preferences of two users (i and j) do not match. The PCC is computed by using the following formula.

$$s_r(i,j) = \frac{\sum_{v \epsilon S_{ij}} (r_{iv} - \overline{r_i})(r_{jv} - \overline{r_j})}{\sqrt{\sum_{v \epsilon S_{ij}} (r_{iv} - \overline{r_u})^2 \sum_{v \epsilon S_{ij}} (r_{jv} - \overline{r_j})^2}}, \tag{4.5}$$

where

$$S_{ij} = \{ v \epsilon V | r_{iv} \neq 0 \wedge r_{jv} \neq 0 \}.$$

In (4.5), the similarity between two experts i and j is calculated only those venues that are visited by both the users.

The similarity calculation in (4.5) results into a very sparse similarity graph because, majority of the venues are not visited by either of the two users. To address the data sparseness problem, we augment the similarity computation with the confidence measure. The confidence measure can be interpreted as a conditional probability that a venue visited by a one user is also visited by the other user in the dataset. The following equation is utilized to calculate the weight of an edge between two users.

$$\omega_{ij} = \begin{cases} s_r(i,j), & if\ s_r(i,j) > 0 \\ & \text{otherwise} \\ P(r_i|r_j) \times \dfrac{1}{1 + \sum_{v \epsilon V_j} |r_{iv} - r_{jv}|} & .P[r_j] \neq 0, \end{cases} \quad (4.6)$$

where, the parameter V_j is the set of venues checked-in by the user j. The parameter P(r_i ⎪ r_j)=P[r_i∩r_j]/P[r_j] is the likelihood ratio that both the user may visit the similar set of venues in future. The additional sum factor in denominator is used to keep value of probability lower than similarity so that the preference must be given to the positive values of similarity. Moreover, in (4.6), if the similarity value is greater than 0, then this value is assigned as an edge weight of the similarity graph. However, when the similarity value is less than zero, then we consider the lower term of (4.6) to assign the edge weight. This implies that an edge is always assigned a non-zero weight that results in the reduction of data sparseness.

The mapping phase also computes the geographical distance of the active user from the popular venues [4.23], [4.24]. The geospatial information about active users and venues are presented as GPS coordinates. Therefore, we utilize Harversine model to compute the user-to-venue distance as follows [4.13]

$$d=Rc, \quad (4.7)$$

In (4.7), the parameter c is the angular distance in radian between current user and venue geospatial location. The parameter R is earth's radius. We use a simple transformation function s_c (u_i,v_j) to calculate the user-to-venue geographical closeness by taking invers of the distance d. The region-wise similarity graph of experienced users and location closeness of popular venues are stored in the database for later online recommendation phase.

4.4.2. Recommendation Module

The recommendation module utilizes bi-objective optimization to generate an optimized list of venues. Suppose an active user A is interested in venue type T that must be located closest to the current location of the active user within a specific region R. In such a scenario, the active user requires the best preferred venues as well as the closest venues from the user's current location. To meet both the aforementioned objectives, we utilize bi-objective optimization in the proposed MobiContext recommendation framework [4.14], [4.8]. The optimization module (see Figure 10) simultaneously maximizes the following two objectives: (a) popular venues and (b) venues' closeness that can be stated as:

$$max\ \ f(o_i)\ \forall\ o_i \in \{\ p_v, v_c\ \}. \tag{4.8}$$

In (4.8), the parameter f(o_i) represents the maximized objective function, in terms of popular venues visited by expert users (〖 p〗 _v) and venue closeness(v_c).

In the subsequent subsection, we discuss in detail the two popular approaches termed as: (a) scalar optimization and (b) vector optimization to address the bi-objective optimization in the MobiContext framework system [4.14], [4.15], [4.19], [4.20], [4.21].

4.4.2.1. Scalar Optimization

In scalar optimization Majority of the classical methods, such as weighted sum and adapted-weighted sum are used to transform multiple objectives into a single aggregate function [4.14]. The aforementioned transformation in BORF is performed using two phases: (a) multiplying each objective function with weights and (b) summing up all the weights and objective functions to generate a single optimal suggestion for a venue. The process of the

abovementioned phases will generate the top-N optimized venue recommendations. The weighted sum approach for BORF can be presented as follows.

$$f(u) = \sum_{i=1}^{n} \alpha_i \times f_i(u), \quad (4.9)$$

In scalar optimization technique, the multiple objectives are transformed into a single-objective aggregate function [4.15].

In (4.9), the function f(u) is the aggregated objective function, the parameter α_i is the weight that determines the significance of n number of objective functions [4.14], [4.20]. In our scenario, there are two objective functions, termed as preferred venue and venue closeness. The weights for preferred venue and venue closeness are formulated in the subsequent text.

4.4.2.1.1. Ccollaborative Filtering-BORF Approach

The proposed CF-BORF utilizes a variant of the CF approach and employs the weighted sum method to implement scalar optimization. The Algorithm 1 illustrates the proposed CF-based approach.

a. Initialization (Line 1):

The algorithm takes the input parameters: (a) active user identification that generates a recommendation query and (b) geographical region where the active user is currently located. Here, active user is the current user connected with the system.

b. Aggregate utility function construction (Line 2–Line 7):

The function computsimset( ) computes the edge weights of the active user with the expert users by utilizing the similarity formula described in (4.5). In Line 5, the function computsimD( ) collects the venues of those expert users that are in the closest proximity of the active user. The aggregate similarity of the active user with the neighbor users is computed in the

Line 6 by utilizing function computeagg( ). The computation of aggregated similarity is performed using the following equation:

$$
\begin{aligned}
s_{agg} &= \frac{\tau}{\tau+\gamma} \times s_r(i,j) + \frac{\gamma}{\tau+\gamma} \times s_c(u_i, v_j), \\
&\text{where} \quad \tau = \frac{s_r(i,j)}{\sum_{k=1}^{n} s_r(i,\ j_k)}, \\
&\text{and} \quad \gamma = \frac{s_c(u_i,\ v_j)}{\sum_{k=1}^{n} s_c(u_i,\ v_k)}, \\
s.t: &\sum_{k=1}^{n} s_r\ (u_i, j_k) \text{ and } \sum_{k=1}^{n} s_c\ (u_i, v_k) \neq 0.
\end{aligned}
\tag{4.10}
$$

In (4.10), the parameter s_agg indicates the overall aggregated similarity with respect to preferred venues and user-to-venue closeness. The user's similarity in terms of preferences is scaled by the average of user's similarity in a specific region by utilizing parameters γ. The user-to-venue closeness is scaled by the average of user-to-venue closeness by utilizing the parameter.

c. Recommendation module (Line 8–Line 9):

On completion of the N number of iterations, the algorithm generates the top-N venues for the user by applying the CF-based recommendation formula stated as follows.

Algorithm 1. CF-BORF-based venue selection

Input: Current User: c, region: R

Output: Top_{rec}= A set S' of top-N venues.

Definitions, *V_e= set of venues visited by expert user e, N_c= set of recommended venues, l_c=location of current user c, V_c = set of venues visited by current user. S_r = set of expert user similar with the current user c, s_d = closeness measure of the expert user e with the location of current user c.*

1: $N_c \leftarrow \emptyset$; $s_{agg} \leftarrow \emptyset$;
2: $S_r \leftarrow computsimset$ (c, E)
3: for $each\ e \in S_r$ do
4: S $\leftarrow \{v: V_e | v \notin V_c\}$
5: $s_d \leftarrow max\big(computsimD(l_c, \mathrm{S})\big)$
6: s_{agg}[e] $\leftarrow computeagg(\ s_e\ , s_d\ )$
7: end for
8: $N_c \leftarrow computRec(\mathrm{c},\ s_{agg})$
9: $Top_{rec} \leftarrow sort\ (N_c)$

Algorithm generates the top-N venues for the user by applying the CF-based recommendation formula stated as follows.

$$r_{cr} = \overline{r_c} + \sum_{e \in S_r} s_{agg}(e, c) \times (r_{ev} - \overline{r_e}), \tag{4.11}$$

4.4.2.1.2. Greedy-BORF Approach

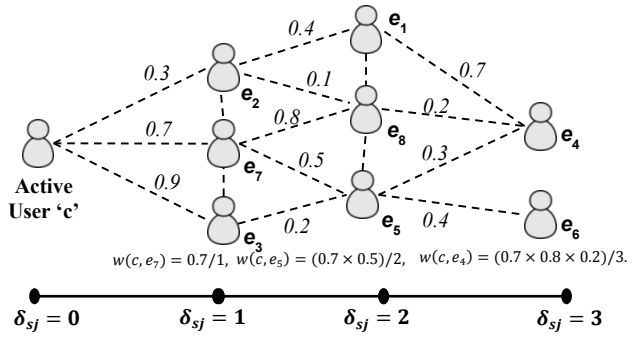

Figure 12. Active user's similarity graph with the experienced users

In this subsection, greedy-BORF approach is presented that generates a set of top-N venue recommendations on a graph of the experienced users. As a first step, the graph of experienced users under a specific region will be retrieved from the dataset. The similarity of the active user will be computed with all of the nodes in the graph using (4.4). New links will be created between the active user and expert users such that their similarity with the active user is greater than zero. The weights of the links are assigned according to the similarity between active user and the expert users. We now refer the experienced users as neighbor nodes. By setting the active user as the root node, immediate neighbors of the active user will be placed at an edge distance of one ($\delta\_sj=1$). Moreover, we select the next active user node among the neighbor nodes as the one that has: (a) maximum weight and (b) maximum number of venues that can be recommended. The neighbor nodes of the newly selected active node will be assigned an edge distance of two ($\delta\_sj=2$). The process continues until the entire graph is traversed or the count of the venues collected from neighbors reaches the value N. Suppose we want to recommend ten venues to the active user. The entries in the table are the check-ins performed by the experienced users at the particular venue as indicated in Table 7 and Figure 12. On the execution of Line 6- Line 10, the venues of the neighbor nodes are collected. The number of venues collected from each neighbor nodes are e3(1), e7(3), and e2(2). The weight of the edge between root node (c) and e3 is greater than root node and e7. However, e3 has less number of venues that can be recommended. We get the following values for each of the neighbor: e_3[$0.9\times1\times(1/10) = 0.09$], e_7[$0.7\times1\times(3/10) = 0.21$], and e_2[$0.3\times1\times(2/10) = 0.06$]. Therefore, on the execution of Line 14, the node e7 will be selected for next level traversal. Line 6- Line 10 will be executed again and the venues collected from the experienced users will be e_1 (0), e_8 (2), and e_5 (3). On the execution of the Line 14 following values will be collected:

e_8[0.8×(1/2)×(2/10)=0.08] and e_5[0.5×(1/2)×(3/10) = 0.075]. Therefore, e_8 will be selected for next level traversal in the graph. As the Line 6-Line 10 are again executed, the venues collected from the neighbors are e_4 (3).The node e_4 does not have any further neighbors. Therefore, the condition of the Line 15 will become true and the execution continues to generate the ranking of the venues on Line 22. Algorithm 2 illustrates the step-by step procedure of the greedy-BORF approach for online recommendations.

Table 7. Number of times required venues are visited by each expert user and total check-ins at the venues

	v_1	v_3	v_4	v_7	v_{11}	v_{12}	v_{22}	v_{25}	v_{44}	v_{45}	$\mathcal{Z}_j$
e_1	-	-	-	-	-	-	-	-	-	-	0
e_2	1	-	-	-	-	-	-	-	8	-	2
e_3	-	7	-	-	-	-	-	-	-	-	1
e_4	53	3	-	9	-	-	-	-	-	-	3
e_5	-	27	13	45	-	-	-	-	-	-	3
e_6	-	-	-	-	-	-	-	41	-	29	2
e_7	-	-	15	-	-	12	16	-	-	-	3
e_8	-	-	-	-	13	-	20	-	-	-	2

a. Initializations (Line 1–Line 4):

The identification of the active user, type of venues to be recommended for active user, and geographical region of active users are taken as the input of the Algorithm 2.

In the Line 2, the similarity graph of the experienced users is retrieved. In Line 3, only those neighbors of active user are selected from the graph that has non-zero similarity computation with the active user. In Line 4, the current user node is stored in the list known as visitedlist.

b. Iterative solution construction (Line 5–Line 22):

In the Line 5, the neighbor nodes (K_a) are sorted in the descending order based on the similarity that is further multiplied by the 1/edge distance between the active user and neighboring node.

Only those venues are selected from the neighboring nodes that were not previously visited by the active user (Line 7).The selected venues are appended in the matrix M. The visited neighbor is stored in the visitedlist (Line 6–Line 10).

If at Line 11, the venue count in the matrix M is greater than the required number of venues N, then the control jumps to Line 22 that computes the geographical distances of the venues in the matrix M from the root node (active user).

c. Aggregate venues provided by the best nodes (Line 23):

The venues are ranked and sorted in the descending order to generate top-N venues to be recommended to the active user. The following equation is used to rank the venues.

In (4.11), x is the venue to be ranked, the parameter c is the active user node, and r_ex is the number of check-ins performed by the expert user e at venue x. The parameter w(c,e) represents the weight of the link in the similarity graph between the root node c and the expert user e. The parameters d(c,x) represents 1/geographical distance between the root node c and the venue x.

4.4.2.2. Vector Optimization

In vector optimization technique, each objective function is presented as a vector. The vector-based approach optimizes all objectives simultaneously in such a manner that a solution cannot improve in one objective without deteriorating the other objective [4.9], [4.14], [4.15]. We design bi-objective vector optimization method termed as GA-BORF by utilizing

evolutionary algorithm (Non-dominated Sorting Genetic Algorithm (NSGA-II)). We selected NSGA-II it proved to be efficient in solving multi-objective optimization problems [4.19], [4.21]. The detailed description of the algorithm is presented below.

Algorithm 2. NSGA-II based venue selection

Input: *R:* set of recommendations.

Output: top-*N* Recommendations based on bi-objective optimization.

Definitions: *Pop= set of population, Epop= set of population after evaluation, gen= number of generations,* Q_t *= Set of top-N optimized recommended venues,* p_{size}*= total size of population.*

1: $parents \leftarrow 0$; $f_L \leftarrow 0$;
2: $Pop \leftarrow randpop\ (p_{size},\ R)$
3: $Epop \leftarrow evaluate\ (Pop)$
4: $PP \leftarrow nondominsort\ (\ Epop)$
5: $S \leftarrow selectParent\ (PP,\ p_{size})$
6: $Q_t \leftarrow crossoverMut\ (S,\ p_{cros},\ p_{mut})$
7: while $(gen \leq max\text{-}gen)$
8: $CC \leftarrow evalute\ (Q_t\ )$
9: $R_t \leftarrow PP\ \cup Q_t$
10: $F \leftarrow nondominsort\ (\ R_t)$
11: for $each\ f_i\ \in F$ do
12: $CDA\ \leftarrow cda\ (f_i\ )$
13: if $size(parent) >\ p_{size}$
14: $f_L\ \leftarrow i$
15: else
16: $parents \leftarrow parents\ \cup\ f_i$
17: end if
18: end for
19: if $size\ (parents) < p_{size}$
20: $f_L\ \leftarrow ccf\ (f_L\ )$
21: $parents \leftarrow parents\ \cup\ f_L$
22: end if
23: $S\ \leftarrow selectParent(parent,\ p_{size})$
24: $pop\ \leftarrow Q_t$
25: $Q_t\ \leftarrow\ crossoverMut\ (S,\ p_{cros},\ p_{mut})$
26: end while
27: return Q_t

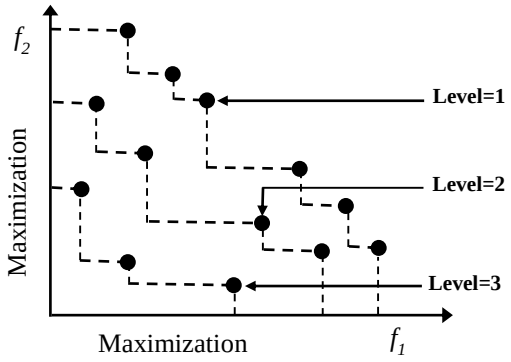

Figure 13. Set of maximized solutions in bi-objective space

The NSGA-II algorithm suggests optimal top-N recommendations and is divided into two phases: (a) recommendation generation and (b) recommendation optimization. The recommendation generation phase uses the CF method with confidence measure as described in the Section 3.1.2 to find out preferred recommended venues. The recommendation optimization phase takes the recommended venues as an input and optimizes based on the preferred location and venue closeness using the NSGA-II [4.19], [4.20]. The NSGA-II presents a set of the candidate solutions called a population. The population of individuals evolves towards the better solutions by employing the genetic operators, such as selection, mutation, and crossover [4.19], [4.20]. In our scenario, each of the individual is defined as a sequence of the top-N recommendation list [r_1,r_2,r_3,…,r_n]. Every single element in the list of recommended venue is termed as a gene. Moreover, every gene in the list (top-N recommendation) consists of: (a) venue identification number and (b) location of the venue (GPS co-ordinates). Algorithm 3 presents the step-by-step description of NSGA-II.

a. Initializations (Line 1– Line 2)

The multiple solutions for a user in the form of suggested recommendations are a list of inputs for the NSGA-II algorithm. In the proposed framework, the recommended venues are

arranged in top-N ascending order. Therefore, we selected permutation-based encoding technique [4.20] to generate population of individuals.

b. Evaluation-based on Objective Functions (Line 3)

In Line 3, the performance of every single individual of the population is evaluated based on the fitness functions.

$$f_1 = \frac{\sum_{i=1}^{n}(ranked\ venues)_i}{t}. \tag{4.12}$$

The fitness function aims to compute the problem specific user defined heuristic [4.20]. The function computes the aggregated ranking score of each recommended venue associated with an individual where the venues' ranks were computed by the HA inference. The fitness function f_1 of an individual in a population is computed as follows:

$$f_2 = \frac{1}{\sum_{i=1}^{n} cost\ (l_u, v)_i \times\ t}. \tag{4.13}$$

In (4.12), the parameter t represents a total number of genes in a single individual. The second fitness function f_2 computes the geospatial distance between the active user's location and the venue of each of the corresponding gene of an individual as follows:

The parameter n represents the total length of an individual. The inverse of the aggregated sum of the cost function cost (v_d,l_u) calculates the geospatial closeness between the current location of the user l_u and the consecutive venues v (genes) of the subsequent ith individuals. The user-to-venue geospatial distance is calculated using Harversine formula described in Section 3.1.2. The fitness function f_2 provides the overall fitness for the venue closeness of a single individual in a population.

c. Selection (Line 4–Line 5)

As specified by NSGA-II, a non-dominated sorting approach is used to classify the entire population. The function nondominsort ( ) acquires an individual (set of recommendations) from the population that is non-dominated from the rest of population. For instance, consider a set of individuals in a population P ={i_1,i_2,...,i_n}. Each individual is assigned fitness functions f_x and f_y. According to non-dominated sorting algorithm, in case of bi-objective optimization, the individual i_c dominates the individual i_(c+1) if and only if:

$$(f_x(i_c) > f_x(i_{c+1})) \quad \text{and} \quad f_y(i_c) \geq f_y(i_{c+1})$$

$$\text{or} \tag{4.14}$$

$$(f_x(i_c) \geq f_x(i_{c+1})) \text{ and } f_y(i_c) > f_y(i_{c+1})$$

According to NSGA-II, all the individuals in a population are sorted based on (4.14). The non-dominated sorting operation is presented in Figure 13. In the case of multi-objective maximization, first level is assigned to the subset of the population that comprises the individuals (set of recommendations) not dominated by any other individuals in the population. Second level is assigned to a subset of the population that presents the individuals not dominated by the remaining unmarked individuals and so on. More precisely, the non-dominated individuals of level zero are the most optimal individuals comprised of the recommendations that are the best suggestions in the trade-off between the venue preferences and venue closeness. Intuitively, the non-dominated set of lowest level will have the highest priority to be a candidate parent for the next population in Line 4.

d. Generate Intermediary Population (Line 6)

The intermediary population is generated by applying the genetic operators, such as crossover and mutation from the set of the parent individuals selected in the Line 6.

In BORF, the sequence of recommendation is important because the top most recommended venue is the highly preferred one for the active user. Therefore, we utilized an ordered crossover method that selects venues from the parent individual as depicted in Figure 14. For instance, let us consider two parents p_1 and p_2 as presented in Figure 14(a). Randomly selected venue-ids are highlighted with the order of visited venues and are copied into ⟦ C⟧_1(see Figure 14 (b)). Moreover, the venues located on the second cut point from parent individual p_2 are copied following the same order as shown in Figure 14 (c).

Mutation operator is a genetic operator that maintains the genetic diversity from one generation of the population of individuals in the next generation [19]. We select the swap mutation operator that is commonly used in a permutation-based representation [20]. Swap mutation generates individuals by randomly swapping two genes from the individual [20].

id	12	21	31	44	53	68	76	88	92
p_1	1	2	**3**	**4**	**5**	6	7	8	9
(a)									
id	88	53	76	12	21	44	92	31	68
p_2	8	5	**7**	**1**	**2**	4	9	3	6
(b)									
id			31	44	53				
C_1	-	-	3	4	5	-	-	-	-
(c)									
id	12	21	31	44	53	92	68	88	76
C_1	1	2	**3**	**4**	**5**	9	6	8	7

Figure 14. Ordered Crossover: (a) and (b) are randomly selected venue-ids, (c) Insertion of randomly selected venue-ids in new offspring C_1 with the same order, and (d) insertion of venues-ids into new offspring from the second cut point of parent $\mathbf{P_2}$

e. Iterative procedure for generating best solutions (Line 7– Line 26)

The Line 7 evaluates the offspring (q_t) depending on the fitness functions described in the (12) and (13). The Line 8 will generate a merged population (R_t) of individual candidates through combining the population of parents and the offspring of size 2N.

The overall population is of size 2N. Therefore, all the individuals that are categorized as a lower-to-higher level using a non-dominated sorted algorithm cannot be accommodated in a new population of size N. To accommodate new size N population (Line 12), the Crowded Distance Assignment (CDA) is calculated. The CDA basically estimates the density of the individuals with respect to the neighboring individuals. In CDA, the average distance of the neighbors of the individuals is calculated to estimate the density of every individual in a specific level. The CDA will be further used in the Crowded Comparison Function (CCF), described in the subsequent text.

To accommodate exactly the N number of population, the individuals that are arranged level-wise are compared. If the number of individuals in a level is less than the total population size of N then the current level will be selected for the next generation (Line 16). However, if the number of individuals in a level is larger than the population size N then the last leveled individuals are organized using the Crowded Comparison Function (CCF) (Line 20). According to the CCF, if the individuals belong to different levels, the individual with the lowest level will be selected for the next generation. Alternatively, if the individuals belong to the same level then the highest crowded distance is selected for the next iteration.

Finally, the best individuals from the merged population undergo a crossover and mutation and are combined with the original population to form a new population of the candidate individuals (Line 23–Line 25). If the number of generations is less than the maximum number of required generations, then the algorithm will perform iteration. Otherwise, the

population of optimal individuals in a form of the top-N recommendations will be generated for a specific region R (Line 27).

4.5. Time Complexity Analysis

In this section, we compute the time complexity of the pre-processing phase, CF-BORF, the greedy-BORF, and GA-BORF approach, respectively.

For a specific number of regions, the time complexity of the HA inference model [16] is $O\ (a\ \times r \times (x'^2 + y^2)$, where the parameter a presents the total number of iterations for approaching to the convergence, x' and y present total number of users and the venues in a region r. The time complexity of the similarity computation for an expert user is $O(r\ \times x^2)$ and the proximity computation graph is $O(r\ \times x \times y)$. The total time complexity of HA, ranking and mapping computation is $O\ (r \times \Big(\big(a \times (x'^2 + y^2)\big) + y \log y\Big)$. For the higher values of venues y, the value of $y\ log\ y$ become insignificant. Therefore, the overall time complexity of the pre-processing phase would be $O\ (r \times \Big(\big(a \times (x'^2 + y^2)\big)\Big)$.

The time complexity of Line 2–Line 7 of the CF-BORF is $O(\ x\ \times\ y^2)$. The Line 8 has an overall complexity of $O\ (x)$. We added all the complexities as $O\ [top_N \times \big((x\ \times\ y^2) + x\big)]$. The value of the top-N is smaller than the total number of expert users and the popular venues.

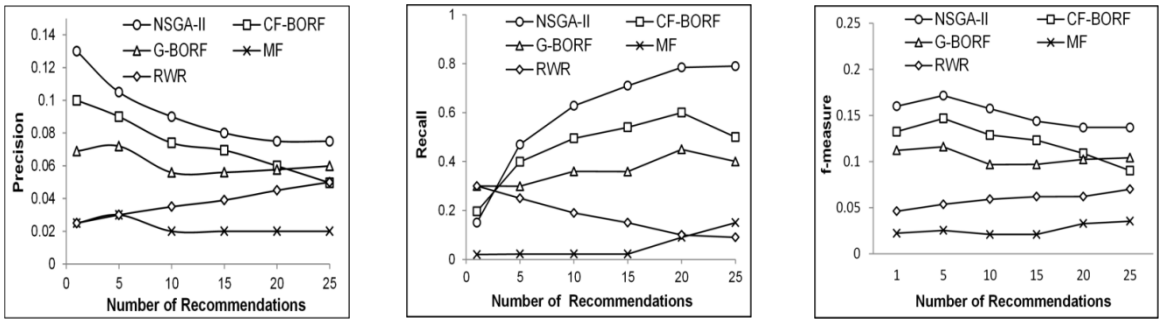

Figure 15. Performance evaluation results: (a) Precision, (b) Recall, and (c) F-measure

Therefore, the overall complexity of CF-based bi-objective optimization algorithm is $O\ (x\ \times y^2)$.

The time complexity of the greedy-BORF computes the similarity with the set of experienced users. The similarity function for y venues is O(y).Therefore, total time complexity of Line 2 for x expert users is O(x+y). The line 5 takes O(x + log x) to sort the x experts. In the worst case, the Line 6- Line 10, number of iterations is x, and the Line 5- Line 14 also takes x iterations. The time complexity of Line 14 is O(x). The combine time complexity of Line 5- Line 14 is O($3y^2$ + x log x). The time complexity of Line 22 is O(n) , where n represents the number of venues that can be recommended and O(x × n) is the time complexity of Line 23. Therefore, the time complexity for Algorithm 1 for 1 region is O(x^2 + x (log x + n)). For r regions, the time complexity becomes O(r(x^2 + x (log x + n))).

In the GA-BORF approach, the time complexity of NSGA-II in Line 1 is $O\ (y^2)$ because of the process of generating the random population of the top-N recommendations of size N in a region. The Line 2 evaluates each individual with respect of objective functions. The time complexity of evaluation function is $O\ (\ M\ (y^2 \times x^2)$, where parameter M is the number of objective functions. To identify the individuals related to a first non-dominated rank in a non-dominated sorting algorithm. Every single individual is compared with the other individual to find the dominance with a complexity of $O\ (\ M\ (y\ \times x))$. For the multiple iterations to find out all the dominated solutions, the total complexity of the non-dominated sorting algorithm is $O\ (\ M\ (y^2 \times x^2)$. The complexity of a crowded distance is $O(x + y\ log\ y)$. We conglomerate the overall time complexity of the NSGA-II-based recommendation algorithm to be as $O\ (x\ log\ \ x)$.

4.6. Performance Evaluation

In this section, we present the performance evaluation of the proposed BORF. We compare our results with following related schemes: (a) User-Based Collaborative Filtering [4.18] (UBCF), (b) Matrix Factorization (MF) [4.17], and (c) Random Walk with Restart (RWR) [4.6]. A brief description of the schemes is presented in the next subsection.

4.6.1. Related Recommendation Techniques

In this section, we discuss approaches commonly used in recommendation systems.

- User-Based Collaborative Filtering (UBCF) computes similar users who visited the similar venues in the past are most likely visit the same venues in the future [4.1], [4.2], [4.3].
- The Matrix Factorization (MF) approach maps the users and venues to a joint latent factor space of a dimensionality a [4.17]. A user x is related to a row vector $\mathbf{p_x} \in \mathbf{R^a}$ and a venue y is associated with a column vector $\mathbf{q_y} \in \mathbf{R^a}$. The estimated rating of the user x for a venue y can be stated as $\mathbf{r_{x,y}} = \mathbf{p_x^T} \times \mathbf{q_y}$, where $\mathbf{r_{x,y}}$ estimates the user's overall interest in a particular venue in VRS.
- The Random Walk with Restart (RWR) method combines the data about frequently visited venues and the friends, represented as social ties in a graph using a structured transition matrix [4.6]. The RWR leverages several sources of the data and encode them into a network structure. The RWR performs a personalized random walk on the graph with a restart to suggest the recommendations for an individual user.

4.6.2. Results

We utilized "Gowalla" dataset consists of 6,442,890 check-ins performed by 150,734 users in total number of 1,280,969 venues [6]. Users with least number of visits have been filtered out by setting the threshold on the number of users' check-ins. The reason for such

filtration is that many of users in Gowalla dataset performed check-ins at very few places that may not add significant contribution into the real-time analysis. We perform extensive experiments on our internal OpenNebula cloud setup running on 96 core Supermicro SuperServer SYS-7047GR-TRF systems. In the selected dataset, out of the entire records, 80% of the record is used as the training set and 20% constitute the test set for the evaluation. We used a standard 5-fold cross validation technique for evaluating the accuracy rate of the framework [4.3].

We utilized the three popular performance evaluation metrics to evaluate the proposed recommendation frameworks: (a) precision, (b) recall, and (c) F-measure. The precision presents a ratio of the accurate recommendations (true positive (tp)) to the total number of anticipated recommendations (tp+ false positive (fp)). An accurate recommendation is the recommendation that has been predicted correctly in the top-N recommended venues.

$$Precision = \frac{tp}{tp + fp}. \tag{4.15}$$

The recall measures the completeness by computing the average quality of the individual recommendations. The recall presents the proportion of all the accurate recommendations in the top-N recommended venues and can be represented as:

$$Recall = \frac{tp}{tp + fn}. \tag{4.16}$$

The F-measure is the harmonic mean of precision and recall and denoted as follows:

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{4.17}$$

As reflected in Figure 15(a), NSGA-II demonstrates the better performance in terms of precision and recall as compared to the rest of the schemes (CF-BORF and greedy-BORF). The

NSGA-II approach optimizes objectives termed as preferred location and location closeness simultaneously, that offers least tradeoff between the objectives. In contrast, the CF-BORF and greedy-BORF approaches present slightly lower performance because of the aggregation method that maps the users' preferences and location closeness into single objective function. Such aggregation cannot provide accurate results especially when there is tradeoff between the user's preferences and location closeness. For instance, in the case of CF-BORF, when there is no similarity between two users' preferred locations, the venue will be suggested to the active user on the bases of user-to-venue closeness. Such suggestion may not provide optimal recommendation and indicates lower performance in terms of precision and recall as presented in Figure 15(a) and 15(b).

We compare the proposed optimization techniques (CF-BORF, greedy-BORF, and GA-BORF) for venue recommendation with the existing UCF, MF, and RWR techniques. As reflected in Figure 15(a), CF-BORF, greedy-BORF, and GA-BORF present the better performance in terms of precision and recall as compared to the rest of the existing schemes, such as UCF, MF, and RWR. The improved performance is because the proposed techniques optimize the recommendation by taking into account the user preferences based on similarity computation and user-venue closeness. The venue suggestions based on such optimization are not only the most preferable for a given user, but also located in the closest proximity of a user's current location. Apart from optimization, CF-BORF, greedy-BORF, and GA-BORF also address the problems of data sparseness by amplifying the similarity computation with the confidence measure that provides precise estimation of user-venue likeness. Moreover, the CF-BORF, greedy-BORF, and GA

BORF methods address the cold start problem by utilizing the HA inference model that helps inferring for the most popular venues within a specific region. The application of confidence measure and HA inference model effectively helps to obtain better solution that results in an increased recommendation precision.

The RWR method demonstrates high performance in terms of precision and recall as compared to the traditional CF-based approach, such as MF and UCF. The reason is that the RWR does not compute the similarity by utilizing user-to-user similarity matrix. Therefore,

RWR is not significantly affected by cold start and data sparseness issues. The UCF indicates very low performance in terms of precision and fails to provide any significant results due to highly sparse dataset of "Gowalla". Therefore, in our experiments UCF is not presented in plots. The tradeoff between precision and recall is depicted in Figure 15(b). Compared to other schemes, the GA-BORF indicates better performance in terms of the F-measure as presented in Figure 15(c).

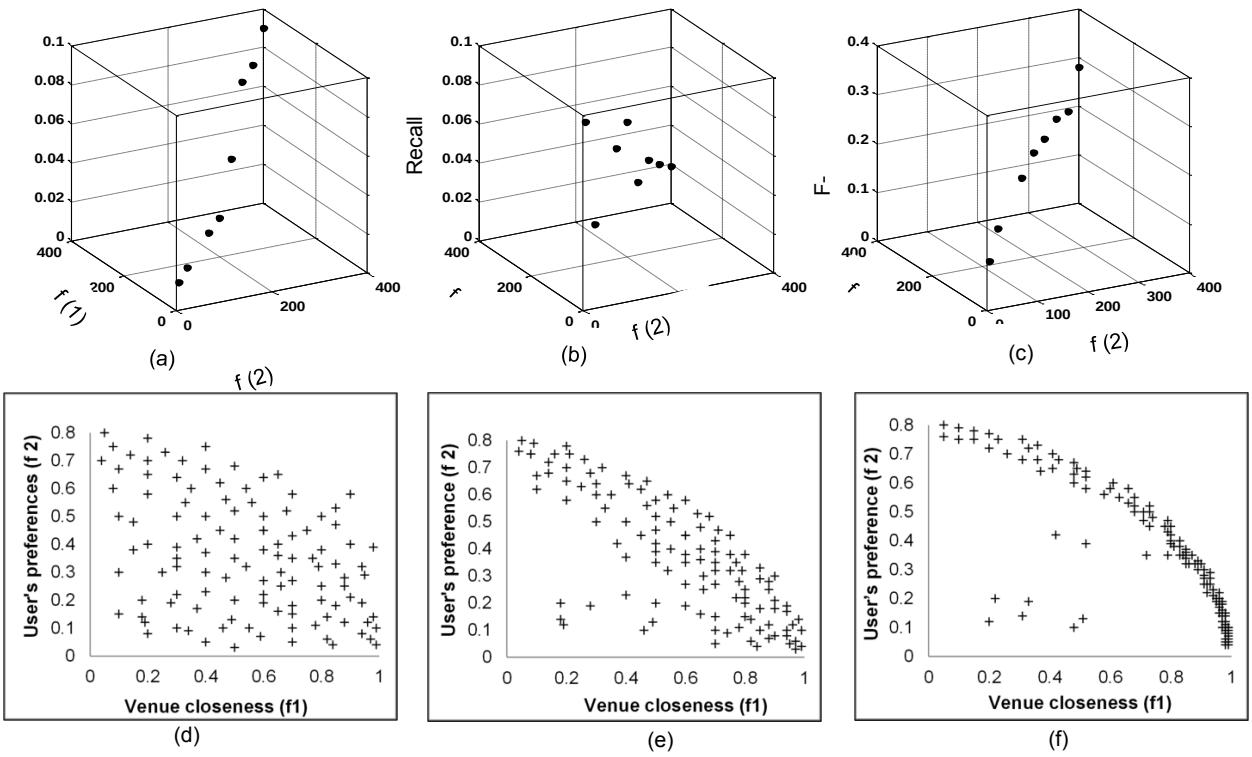

Figure 16. Multi-objective performance measure: (a) Precision, (b) Recall, (c) F-measure for NSGA-II, (d) Generation size 5, (e) Generation size 100, and (f) Generation size 200

Figure 16presents the bi-objective performance measure in terms of precision, recall, and F-measure. Users' preferences and venues' closeness as bi-objectives in NSGA-II demonstrates better performance as we increase the number of generations. Figure 16(a) presents increase in precision. Alternatively, Figure 16(b) and 16(c) indicate improved performance in term of recall and F-measure.

A series of the simulation runs were conducted to test the effectiveness of the of NSGA-II algorithm. The NSGA- II reports best performance for crossover rate = 0 .9 and mutation rate = 0.1, respectively. These parameter values were determined empirically through numerous runs on "Gowalla" datasets. The Figure 16(d), 16(e), and 16(f) present the population at the generation 5, 100, and 200, respectively. Moreover, the convergence of population solutions toward the

optimization of both objectives is clear. The Pareto front [4.20] depicted by Figure 16(f) contains the best solutions with regard the two objectives (users' preferences and location closeness).

4.7. Related Work

In the past, most work focused on mining geographical locations of user and predicting the user's movement among these locations [4.7], [4.10], [4.22]. Recently, many scientific literature [4.6], [4.7] have highlighted the significant impact of incorporating the user's geospatial information with cloud infrastructure into traditional VRSs. Some of the latest efforts have been highlighted in the subsequent text. An on demand ubiquitous cloud-based venue recommendation system was proposed by Khalid et al. The authors utilize HITS method, Ant colony optimization, and collaborative filtering on a cloud infrastructure to provide optimal venue recommendations. Similarly, a personalized venue recommendation system is introduced by Zheng et al. [4.10] that provides interesting venues and classical correlation between the users' travel experiences mind from GPS trajectories. The system provides venue recommendations to the user based on the user's travel sequences [4.10]. A similar approach is presented in [4.3] and [4.11] that keep track of users' traveling history, reduce computation cost, and suggest the best route to the user by utilizing the cloud infrastructure. Differing form aforementioned work, we do not keep track of user's travel history that causes heavy computation. In our proposed MobiContext BORF, we mine the similarity between users and proximity of interesting venues that predicts the user's interests and preferences for an unvisited venue.

Apart from recommending the best routes by observing users' past traveling experiences, few existing recommendation techniques have based their model on implicit ratings. Implicit rating presents number of visits (check-ins) performed by user at different places. For instance,

Hsun-Ping Hsies et al. [4.7] proposed a recommendation technique to suggest time-sensitive trip routes based on the user's check-in performed at different geographical regions. Similarly, the author in [4.6] proposed a model based on random-walk-with-restart approach that suggests venue recommendations by acquiring users-venue-check-in information. Majority of the aforementioned recommendation techniques provide relevant suggestions about venues. However, such techniques suffer from problems, such as cold start and data sparseness that occur when a user has checked-in at only limited number of venues out of many existing in database. In such case, there is sparsely filled user to venue check-in matrix that may yield to the loss of recommendation accuracies. Our proposed MobiContext BORF framework addresses these issues and presents solutions for data sparseness and cold star. Moreover, we specifically incorporate bi-objective optimization techniques in the proposed framework to optimize users' preferences, and location closeness from the venues.

4.8. Conclusions

We have presented a multifold contribution by contriving a cloud-based bi-objective optimized solution MobiContext Bi-Objective Recommendation Framework (BORF) that produces optimized recommendations by simultaneously considering the trade-offs among real-world physical factors, such as person's geographical location, distance of person from venue, and travel conditions. The significance and novelty of the proposed framework is the adaptation of popular collaborative filtering and bi-objective optimization approaches, such as scalar and vector. The venue suggested by the proposed BORF is not only the most preferable suggestion for a user, but also located in the closest proximity of a user's current location. In our proposed approach, data sparseness issue is address by integrating the user-to-user similarity computation with confidence measure that quantifies the amount of similar interest indicated by the two users

in the venues commonly visited by both of them. Moreover, cold start issue is resolves by introducing the HA inference model that assigns ranking to the users and venues based on a mutual reinforcement relationship. In such case, the BORF always has a precompiled set of popular unvisited venues that can be recommended to the new user.

The proposed MobiContext BORF implements a variant of collaborative filtering, greedy, and NSGA-II based algorithms for the optimized venue recommendation. The evaluation results on a real-world dataset "Gowalla" indicate that NSGA-II based venue recommendation approach outperform most of the existing VRSs.

In the future, we would like to extend our work by incorporating more contextual information in the form of objective functions, such as the check-in time, users' profiles, and interests, in our proposed framework. Moreover, we intend to integrate other approaches, such as machine learning, text mining, and artificial neural networks to refine our existing framework.

4.9. References

[4.1] A. Majid, L. Chen, G. Chen, H. Turab, I. Hussain, and J. Woodward, "A Context-aware Personalized Travel Recommendation System based on Geo-tagged Social Media Data Mining," International Journal of Geographical Information Science, pp. 662-684, 2013.

[4.2] C. Chow, J. BAO, and M. Mokbel, "Towards Location-Based Social Networking Services," In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, ACM, pp. 31-38, 2010.

[4.3] J. Levandoski, M. Sarwat, A. Eldawy, and M. Mokbel, "Lars: A Location-aware Recommender System," In IEEE 28th International Conference on Data Engineering (ICDE), pp. 450-461, 2012.

[4.4] F. Liu, and H.J. Lee, "Use of Social Network Information to Enhance Collaborative Filtering Performance," Expert System with Applications, vol. 37, no. 7, pp. 4772-4778, Jul. 2010.

[4.5] J. Bobadilla, F. Ortega, A. Hernando, A. Gutierrez, "Recommender Systems Survey," Knowledge-Based Systems, vol. 46, pp. 109-132, July 2013.

[4.6] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "A Random Walk around the City: New Venue Recommendation in Location-Based Social Networks," In Proceedings of International Conference on Social Computing (SocialCom), IEEE, pp.144-153, 2012.

[4.7] H. P. Hsieh, C. Te, S. Lin, "Exploiting Large-Scale Check-in Data to Recommend Time Sensitive Route," In Proceedings of 12th International Conference of Urban Computing, pp. 55-62, 2012.

[4.8] J. Bao, Y. Zheng, M. Mokbel, "Location-based and Preference-Aware Recommendation Using Sparse Geo-Social Networking Data," In Proceedings of ACM SIGSPATIAL GIS, 2012.

[4.9] A. Said, B. Jain, S. Albayrak, " A 3D Approach to Recommender System evaluation," In Proceeding of 13th International Conference of Computer Supported cooperative Work Companion (CSCW), pp. 263-266.

[4.10] O. Khalid, M. U. S. Khan, S. U. Khan, and A. y. Zomaya, "OmniSuggest: A Ubiquitous Cloud based Context aware Recommendation system for Mobile Social Networks," IEEE Transaction on Services Computing, 2013.

[4.11] J. J. Ching, E. H. Chan, B. Shi, and V. Tseng, "TripCloud: An Intelligent Cloud-based Trip Recommendation System,"In Proceedings of International Conference on Advances in Spatial and Temporal Databases, pp. 234-256, 2013.

[4.12] Y. Wang, S. Wang, N. Stash, L. Aroyo, and G. Schreiber, "Enhancing Content-Based Recommendation with the Task Model of Classification," In Proceedings of the Knowledge and Management, pp. 431-440, 2010.

[4.13] S. Seema, and S. Alex, "Dynamic Bus Arrival Time Prediction, using GPS Data," In Proceedings of the Nat. Conf. Technological Trends (NCTT), Nov. 2010.

[4.14] T. Jambor, and J. Wang, "Optimizing Multiple Objectives in Collaborative Filtering," In Proceedings of 4th ACM Conference on Recommender System, pp. 55-62. 2010.

[4.15] M. Ribeiro, A. Lacerda, A. Veloso, and N. Ziviani, "Pareto-Efficient Hybridization for Multi-objective Recommender Systems," In Proceeding of 6th ACM Conference on Recommender Systems, pp. 19-26, 2012.

[4.16] B. Chandra, S. Bhaskar, "Patterned Growth Algorithm using Hub-Averaging without Pre-assigned Weights," In Proceeding of IEEE International Conference on Systems, man, and Cybernetics (SMC), pp.3518-3523, 2010.

[4.17] B. Hidasi, and D. Tikk, "Initializing Matrix Factorization Methods on Implicit Feedback Database," Journal of Universal Computer Science, vol. 19, no. 12, pp. 1835-1853.

[4.18] Q. Qi, Z. Chen, J. Liu, C. Hui, and Q. Wu, "Using Inferred tag rating to Improve User-based Collaborative Filtering," In Proceedings of the 27th ACM Symposium on Applied Computing, ACM, pp. 2008-2013, 2012.

[4.19] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast and Elitist Multi-objective Genetic Algorithm: NSGA-II," IEEE Transaction on Evolutionary Computations, vol. 6, no. 2, pp. 182-197, 2002.

[4.20] C. Chitra and P. Subbaraj, "A Non-dominated Sorting Genetic Algorithm for Shortest Path Routing Problem," International Journal of Electrical and Computer Engineering, 2010.

[4.21] A. Said, B. Jain, S. Albayrak, " A 3D Approach to Recommender System evaluation," In Proceeding of 13th International Conference of Computer Supported cooperative Work Companion (CSCW), pp. 263-266.

[4.22] M. Ye, P. Yin, W. Lee, D. Lee, "Exploiting Geographical Influence for Collaborative Point-of-Interest Recommendation," In Proceeding of 34th International Conference on research and Development in Information Retrieval, pp. 325-334, 2011.

[4.23] Q. Yuan, G. Cong, Z. Ma, and A. Sun, " Time-Aware point-of-Interest Recommendation," In Proceedings of 36th International Conference on Research and Development in Information Retrieval, pp. 363-372, 2013.

[4.24] J. Zhang, C. Chow, "iGSLR: Personalized Geo-Social Location Recommendation: A Kernel Density Estimation Approach", In Proceedings of 13th International Conference on Advances in Geographic Information System, pp. 334-343, 2013.

5. SOCIALREC: A CONTEXT-AWARE RECOMMENDATION FRAMEWORK WITH EXPLICIT SEMENTIC ANALYSIS

5.1. Abstract

In recent years, recommendation systems have seen significant evolution in the field of knowledge engineering. Most of the existing recommendation systems based their models on collaborative filtering approaches that make them simple to implement. However, performance of most of the existing collaborative filtering-based recommendation system suffers due to the challenges, such as: (a) cold start, (b) data sparseness, and (c) scalability. In this paper, we proposed a SocialRec, a context-aware recommendation framework that utilizes a rating inference approach to incorporate textual users' review into traditional collaborative filtering methods for personalized recommendations. To address the issues pertaining to cold start and data sparseness the SocialRec utilizes the textual reviews as an additional source of user preference. Moreover, the SocialRec performs preprocessing by using the Hub-Average (HA) inference model. The results of comprehensive experiments on a large-scale real dataset confirm the accuracy of the proposed recommendation framework.

Index Terms—, Multi-objective optimization, Collaborative Filtering (CF), Non-dominated Sorting Genetic Algorithm (NSGA-II).

5.2. Introduction

The advancement in communication infrastructure and easy access to information available through Web has shifted the researchers' attention from information acquisition problem to information retrieval problem. Moreover, the Web users are not only consuming, but also contributing and disseminating information in a vastly decentralized manner via social networks, such as sharing reviews and personal interests [5.1]. The continuous accumulation of

such massive Web contents consequently leads to the problem of information overload. An autonomous information retrieval system, termed as recommendation systems have become a promising research area as a response to the information overload in recent decade [5.2].

5.2.1. Research Motivation

Recommendation systems are increasingly emerging as an integral component of e-business applications [5.1]. For instance, the integrated recommendation system of Amazon provides personalized recommendations for various items of interest to customers. Recommendation systems utilize various knowledge discovery techniques on a user's historical data and current context to recommend products and services that best match the user's preferences [5.1], [5.2].

In recent years, emergence of numerous social networking services, such as, Facebook, Google Latitude, and Yelp has significantly gained the attraction of a large number of subscribers [5.2], [5.6]. Such social networking services not only allow user to provide explicit feedback in a form of preference rating (star rating), but also allow users to provide textual review about the venue visited by the user [5.2], [5.3]. The large number of such feedback on daily bases results in the accumulation of massive volumes of data. Based on the data stored by such services, several Venue-based Recommendation Systems (VRS) were developed [5.1]–[5.4]. Such systems are designed to perform recommendation of venues to users that most closely match with users' preferences. Despite having very promising features, the VRS suffer with numerous limitations and challenges. A major research challenge for such systems is to process data at the real-time and extract preferred venues from a massively huge and diverse dataset of users' historical feedbacks [5.1], [5.3], [5.4].

Another promising challenge of preference rating is the inherent biasness caused by users' personal interest, choice and current trends in the form of social influence. Such biasness can contaminate the recommender system's performance in terms of accuracy and precision; consequently weaken the system's ability to provide high-quality recommendations. For instance, Figure 17 demonstrates the biasness in the comments from two different reviews for a restaurant. There is a clear difference of opinion in rating assignments by the two reviewers. As presented in the figure, the two users (Michelle and Clif) wrote about quality of a restaurant in LA, USA. Both the users have provided a positive feedback and seem to be pleased with their experience at the restaurant. The users described the restaurant service with multiple positive words, such as "perfection", "great experience", "awesome". However, the first user gave 5 stars to the restaurant whereas the second user gave three stars.

User-generated preference rating can play a significant role in the popularity of a venue. However, that rating systems are often targeted by rating spammers who seek to distort the perceived popularity of a venue by creating fraudulent rating. To improve the popularity of any particular venue, one of the business tricks is to hire people that make fake identify and rate the desired venue high by assigning highest star rating. Such rating will increase the overall popularity of the venue and the targeted venue becomes the popular place amongst the other venues that do not present the actual popularity trend. All the above-mentioned anomalies become our motivation to improve the existing recommendation systems.

5.2.2. Research Problem

In scientific literature, several works, such as [5.1]–[5.6], and [10] have applied Collaborative Filtering (CF) to the VRS recommendation problem. The CF-based approaches in tend to generate recommendations based on the similarity in actions and preferences of users

[5.3], [5.2], [5.5]. Generally, in CF-based recommendation systems acquire user preferences through explicit feedback that is obtained by directly querying the user. In such systems, users are presented with preference integer rating (5 star rating) to quantify the preferences. However, many users prefer to use free form of text to express there opinion. For instance, reviewers written by travelers on tourism destinations are popular source of information that influences the users' choice of destination. Moreover, one may want to extract information apart from preference integer rating. For instance, a user wants to acquire information about a certain feature or aspect of a venue, such as quality of service and decor. Despite the importance and value of such information, there is no comprehensive mechanism that formalizes the opinion selection and retrieval process. In such scenarios, CF-based recommendation systems ignore the significance of feedback that is embedded in the textual review especially in a scenario when enough explicit feedback is not available in the form of integer scale [5.4]. Therefore, it is consequential to bridge the gap between opinion mining from textual data and recommendation systems and to go beyond the information conveyed by the preference integer rating by utilizing the free-text user review for recommendation.

Despite being less complicated, most CF-based recommendation techniques suffer from several limitations that make them less ideal choice in many real-life practical applications [5.2]. The following are the most common factors that affect the performance of many existing CF-based recommendation systems:

- Cold start. The cold start problem occurs when a recommendation system has to suggest venues to the user that is newer to the system [5.3]. Insufficient check-ins for the new user results in zero similarity that degrades the performance of the recommendation system [5.2].

The only way for the system to provide recommendation in such scenario is to wait for sufficient check-ins by the user at different venues.

- Data sparseness. Many existing recommendation systems suffer from data sparseness problem that occurs when users have visited only a limited number of venues [5.4]. This results into a sparsely filled user-to-venue preference matrix. The sparseness of such matrix creates difficulty in finding sufficient reliable similar users to generate good quality recommendation.

Figure 17. Biasness in user's rating

Scalability. Majority of the traditional recommendation systems suffer from scalability issues. The fast and dynamic expansion of number of users causes recommender system to parse millions of check-in records to find the set of similar users. Some of the recommendation systems [5.3], [5.4] employ data mining and machine learning techniques to reduce the dataset size. However, there is an inherent tradeoff between reduced dataset size and recommendation quality [5.1].

- Rating biasness. Inconsistent rating leads to rating biasness resulting in potential noise

- Textual feedback. Most of the existing CF-based recommendation systems are designed based on explicit preference rating. However, in the case of textual feedback in the form of reviews, CF-based recommendation systems cannot recommend the venue to a user.
- The immediate effect of the abovementioned issues is the degradation in authenticity of most of the CF-based recommendation systems. Therefore, it is not adequate to rely solely on simplistic but memory-intensive CF approach to generate recommendations.

5.2.3. Methods and Contributions

In this paper, we propose a context-aware recommendation that overcomes the limitations exhibited by CF-based approaches. To address the cold start issues, our framework utilizes the Hub-Average (HA) inference model [16] that maintains a pre-computed list of most popular venues in a user's current vicinity. To address data sparseness caused by zero values of similarities, we enhanced CF algorithm by utilizing textual review as an additional source of user preferences. The extended version of CF enables the system to recommend where the preference are too complex to be expressed as scalar rating.

In summary, the contributions of our work are as follows.

- We proposed a Context-aware recommendation framework SocialRec that utilized the aggregated preference score acquired from preference rating and textual opinion to suggest optimal venues recommendations.
- Several data cleansing procedure are performed on the extracted data to minimize the data sparcity and cold start problem.
- We perform extensive experiments on our internal OpenNebula cloud setup running on 96 core Supermicro SuperServer SYS-7047GR-TRF systems. The experiments were conducted on real-world "Yelp" dataset [4].

The rest of the paper is organized as follows. Section 2 presents the system overview. In Section 3, we discuss the SocialRec. Section 4 presents the complexity analysis of the proposed framework. In Section 5, we present the performance evaluation with simulation results. The related work is reviewed in Section 6, and Section 7 concludes the paper.

5.3. System Overview

In the subsequent text we will explain the decentralized SocialRec for efficient, fast, and optimized venue recommendation in detail. The proposed system simultaneously considers a user's preferences, textual contents, and past venue preference score when generating online recommendations.

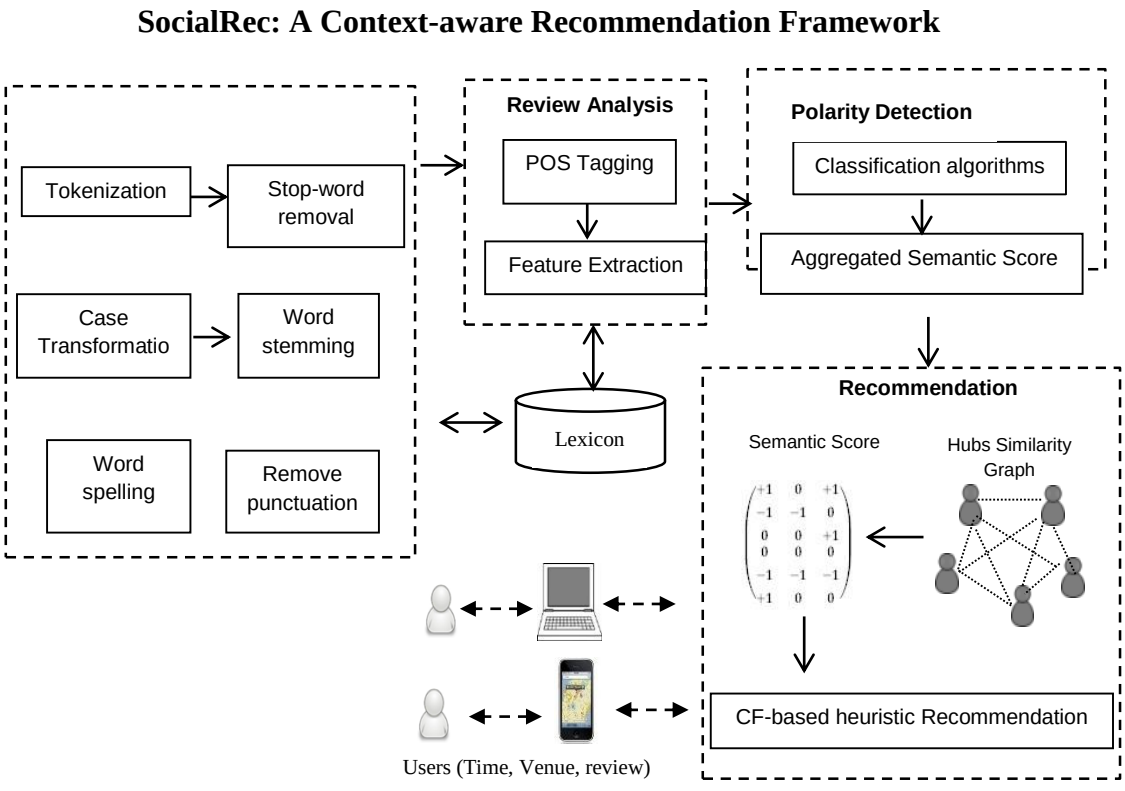

Figure 18. Polarity detection with machine learning algorithms

5.3.1. Major Components

The SocialRec maintains a preference rating for each reviewer by detecting polarity from the textual review. Moreover, preference rating history also maintains a record of the sets of venues preference rating, the user's identification, venues' names commented by the user.

To obtain venue recommendations, a user communicates with the framework through recommendation request queries that consist of: (a) current time, (b) venue type, such as restaurant, café, and bar.

As presented in Figure 18, the proposed framework architecture comprises of four modules, namely: (a) review preprocessing module, (b) review analysis module (c) polarity detection module, and (d) recommendation module.

The preprocessing module transform the unstructured textual data into tokenized and structured format by eliminating noisy text, such as spelling mistakes, grammatical errors, and improper casing [5.7]. Preprocessing module ensures the quality of the text in terms of comprehensibility and representativeness various steps, such as tokenization, word stemming, stop-word removal have been implemented to refine the text for further offline processing. Moreover, we define the boundaries of the text for sentence-wise better understanding. All the above-mentioned pre-processing phases refine the data that will be aggregated and utilized during the next offline review analysis module.

Review analysis module analyzes each and every sentence of the review and categorizes the words in each sentence according to the grammatical structure. POS tagging technique have been implemented to categorize the words of the sentence syntactically that play a vital role in identification of relevant feature and opinion of the comment generated by a reviewer. Moreover,

we proposed a method to extract the relevant features that are mostly under discussion in feature extraction process.

In Polarity detection phase we compute the polarity of every single sentence in a review using different classification algorithms, such as Naïve bayes and Support Vector Machine (SVM). Such classification algorithms assign a polarity score to every sentence in a review. Finally, we calculated the aggregated polarity score of every review that can be further utilize for the online recommendation process.

Online recommendation module inputs the semantic score of the review and recommends the top-N venues for an active reviewer (N is the number of venues recommended by the framework). The recommendation module computes numerical ranks for reviewers and commented venues by utilizing the HA-based inference model [5.6]. The basic idea of the HA-based inference model is to assign ranking to the reviewers and venues based on a mutual reinforcement relationship [5.6]. An expert reviewer is defined as the one who has commented many higher score venues, and is assigned a higher rank by HA. Similarly, a venue that is commented by many highly ranked reviewers gets a higher score and is known as a popular venue [5.15], [5.16]. Moreover, the recommendation module extracts a similarity graph of the experienced reviewers. The reviewer and venues that have very low scores are pruned from the dataset during online recommendation phase to reduce the online processing time. The recommendation module utilizes the CF-based heuristic approach to generate suggestions in the form of venues that best matches reviewers' preferences. The venues at the top of the recommended list will be the ones that most satisfy the reviewers' preferences.

5.4. SocialRec: A Context-aware Recommendation Framework

In this section, we discuss in detail the proposed SocialRec, a context-aware Recommendation Framework. The framework has four main components: (a) Review pre-processing, (b) Review Analysis, (c) Polarity detection, and (d) Recommendation. The detailed description of the above mentioned components is presented in the following subsequent sections.

5.4.1. Review Pre-processing

User-generated online reviews are a short informal text written by visitors require preprocessing phase to remove noisy text, such as grammatical mistakes, spelling errors, improper casing, ad-hoc abbreviations, incorrect punctuations, and malformed sentences [5.9]. Such noise in informal text cause more complications to the mining process and make the dimensionality of the text high, because every single word in the text is treated as one dimension. Therefore, to reduce the dimensionality in text and to improve the performance of mining process, pre-processing phase is the crucial task to perform. Scientific literature witnessed the implication of pre-processing phase as a substantial improvement of text classification process [5.9]. Four common preprocessing steps include word stemming, tokenization, POS-tagging, stop-word removal, lowercase conversions are considered within the scope of this paper.

5.4.1.1. Word Stemming, Tokenization, Stop-word Removal

Tokenization is the process of splitting the sentence into different words, such as number, punctuation marks, and names [11]. Another morphological technique is remove-stop-word. Stop-words, such as “the”, “am”, “an”, and “a”, construct the syntactic structure of the sentence and are the most frequently occurring words. However, these words do not contribute enough to represent the information [5.9], [5.11]. Therefore, stop-words are removed from the text corpus.

Word stemming is another morphological technique that refers to a linguistic normalization to remove the prefixes and suffixes from a word. For instance, the word “connection” is reduced to the root word “connect.”

5.4.1.2. Lower to Upper Case Transformation

Proper use of lower case and upper case is necessary for the syntactic interpretation of the sentence. Syntactically correct sentences end with predefined punctuation markers, such as exclamation mark (!), full stop (.) and interrogation mark (?). We have employed rule-based approach to identify sentence boundaries in noisy text. Sentence boundary detection comprises of two major tasks: (a) identifying end of sentences based on correct punctuation mark implication and (b) disambiguation of full stop (.) from decimal point and abbreviated ending. Basic rules for sentence boundary detection are presented in algorithm 3. In Line (2-10) if the word end with a symbol “.” and the word is not preceded by a pre-defined set of words defined in the dictionary, such as Org., Prof. then the symbol “.” can be treated as a sentences boundary. The Alphabet immediately after the sentence boundary can be converted into uppercase letter. Moreover, in Line 11, the symbol “.” will be ignored if it appears immediately after or before the digit.

5.4.1.3. Irrational Use of Punctuation Marks

Irrational use of punctuation also causes noise in the text. If a punctuation symbol is a valid mark and at the end of the sentence than only one instance of the symbol will be retained in the sentence. Similarly, if the symbol is not a valid punctuation mark than the symbol can be remove from the sentence. For instance a visitor posted a reviews may read as follows:

Example: Grape Leaves are a popular starter. An order yields a little collection of cigar shaped rolls with the perfect ratio of soft, supple leaves and flavorful rice will become Grape

Leaves are a popular starter. An order yields a little collection of cigar shaped rolls with the perfect ratio of soft, supple leaves and flavorful rice!

5.4.1.4. Word Spelling

Erroneous spellings are the major hindrance to extract meaning from text. If a word is erroneously spelt then it may leads to incorrect interpretation of the meaning associated with a text. We have employed a PyEnchant library [5.12], a free available spell checker that replaces the miss-spelt word with the most probable correct word from the dictionary. It is worth mentioning here that the focus of the system is not to correct all the errors in the text. The basic purpose of preprocessing phase is to focus on minimizing errors in opinion mining.

5.4.2. Review Analysis

The review Analysis phase analyzes the linguistic features of review so that the opinion about the review can be identified. Tow majorly adopted task for review analysis are POS (Part-Of-Speech) tagging and Feature Extraction. The detailed description of the aforementioned is presented in subsequent text.

5.4.2.1. POS Tagging

Part-of-speech tagging is the important step in the framework. POS tagging reflects the syntactic category of the word that play a vital role in identification of relevant feature, opinion from reviewer sentences. A rule-based approach (Brill tagging) is implemented using nltk [5.13] to parse each review and split text into sentences. Such sentences are further divided and assigned part-of speech tag for each single word. The taggers extract the noun, verbs, and adjective information from the reviewer's comments. The review sentence with the POS tag is further use for feature extraction (subjectivity detection) and feature reduction steps.

Algorithm 3. Sentence boundaries transformation

Input: A Set R of Review

Output: A set R' of bounded sentences in a review set *R*.

Definitions: *{D} = set of pre-define words in a dictionary, w = set of words in a review.*

1: for $each\ word\ w\ \in$ R do
2: if $[w.end] ==$ "." and $[w.end - 1] \neq \{D\}$
3: $[w.end + 1] \leftarrow uppercase$
4: else
5: $no\ \ boundry$
6: end if
7: if $[w.end] ==$ "."
8 $[w.end + 1] \leftarrow uppercase$
9: end if
8: if $[w.end] ==$ "." and
9: $[\mathrm{w.end} + 1]\mathrm{or}\ [\mathrm{w.end} - 1] \leftarrow\ \ digit$
10: $no\ \ boundry$
11: end if
12: end for
13 return R'

5.4.2.2. Feature Extraction and Reduction

Sentence level sentiment classification comprises of feature extraction and feature reduction phases. Feature extraction process identifies subjective and objective sentences from the review. Objective sentence in a review do not contain users' opinion, whereas subjective sentences contain users' opinion. For instance sentence 1 is objective sentence and sentence 2 and 3 are subjective in the following review

(1) Me and my friend visited Sam Choy's restaurant. (2) Breakfast sandwich served on French toast with a side of syrup is great. (3) For lunch, their hot Panini sandwiches are excellent.

Scientific literature reveals as a good indicator of opinion. Moreover, noun (life, Help, issue, pain), verb (like, degrade), and adjective (bitter, delicious) are used for subjectivity determination of a word. We used SentiWord Net, a lexical resource specifically designs for sentiment classification and opinion mining applications. A single sentence in a review if comprises of noun, verb, and adjective are referred as subjective sentence, otherwise the sentence is referred as objective. The objective sentences do not contribute enough in opinion orientation.

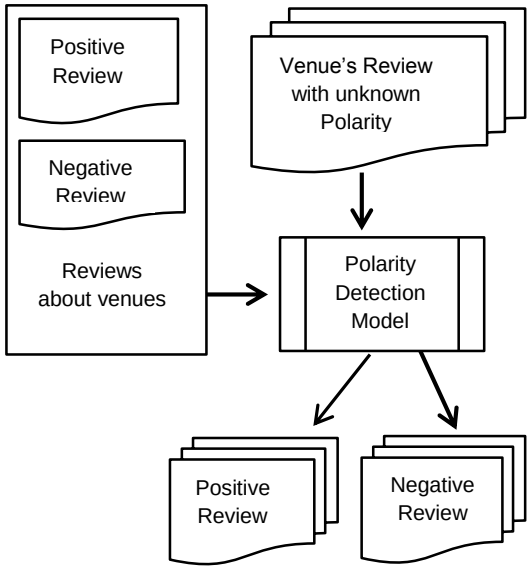

Figure 19. Polarity detection with machine learning algorithms

Therefore, the objective sentences are extracted from the review in the phase of feature reduction. Feature reduction step reduce the dimensionally of the reviewed comment. Consequently reveal better results in classification process.

The detail description of feature-opinion extraction process is described in Algorithm 4. In most of the sentences opinions are expressed by utilizing objectives. A set of collection of reviews are the input to the algorithm. We extracted the opinion from every single sentence s of each Review R. The opinions are obtained by extracting the adjective of every sentence in Line

4. Simultaneously, we assign the Opinion-Feature pair (OFP) as (reviewer_id, venue_id, sentence_id, ∅, adj) by keeping the feature empty into a global OFP set in Line 6. In Line 3- Line 7, with the initial opinion set extracted based on the adjective, we will extract the features associated with every opinion in Line 8- Line 15. Features are usually presented as noun or noun phase. In a single sentence, each noun centered within an opinion window is added to the candidate feature set and the corresponding OFP are updated as. We assume the distance between two neighbor words is 1. For the opinion window in a sentence, we assume the opinion as a center point. Each noun or noun phrase with the distance to the center less than 5 is extracted by utilizing ExtractNoun ( ) function in Line 10. Consequently, the frequency of every noun or noun phrase (if repeated) is also accumulated in Line 11. The OFP is also updated with the newly derived noun or noun phrase as (reviewer_id, venue_id, sentence_id, feature, adj, c). In Line 16- Line 19.only those pair of OFP is extracted that have higher noun frequency count than user defined threshold frequency. We are more interested in the frequently discussed feature therefor the infrequently discussed features are prune and the updated list of OFP will be further used for polarity detection using classification.

5.4.3. Polarity Detection

Polarity detection process classifies sentences of a review as positive, negative, and neutral. We prefer sentence-level polarity classification as the sentence-level polarity detection provides a more fine-grained interpretation of each sentence in a review [5.14]. The basic mechanism of polarity detection by classification is presented in Figure 18.

Various classification algorithms have been widely deployed for polarity detection [5.8]. Each review has a preference integer rating (1-5 stars). Therefor training and testing data are readily available. Generally, a review with 4-5 preference rating is considered a positive review

and a review with preference rating 1-2 is considered a negative review. Scientific literature present various supervised learning methods to classify the sentences into positive, negative and neutral sentences. We selected Naïve Bayes and Support Vector Machine (SVM) for the classification of the sentences into positive, negative, and neutral sentences. Both techniques outperform in text classification [5.8].

Algorithm 4. Feature-opinion extraction

Input: A collection of n number of reviews R = { r_1 , r_2 , r_3 , … , r_n },

Output: FPlist = A set of opinion-feature pair.

1: tempF ← ∅; tempAdj← ∅; c ← 0
3: for $each\ sentence\ s\ in\ Review\ R$ do
4: *P← ExtractOpenion (s, adj)*
5: $tempAdj_s$ ←*P*
6: FOP_s ← *(user_id, venue_id, s_id, ∅ ,* $tempAdj_s$*)*
7: end for
8: for $each\ sentence\ s\ in\ Review\ R$ do
9: for $each\ adj\ in\ tempAdj_s$
10: $nph \leftarrow ExtractNoun\ (s, adj)$
11: $c \leftarrow noun\_frequen\_count\ (nph)$
12: *tempF←* nph
13: FOP_s ← *(user_id, venue_id, s_id,* $tempF$*, , c)*$tempAdj_s$
14: end for
15: end for
16: for each sentence s *in Review R*
17: if $FOP_s.c$ > *threshold frequency*
18: $FPlist \leftarrow FPlist \cup \{FOP_s\}$
19: end for
20: return *FPlist*

5.4.3.1. Naïve Bayes Model for Sentiment Classification

Naïve Bayes classifier is a probabilistic machine learning technique for the text classification [5.15]. The classifier models the distribution of the reviews in each class using a

probabilistic model. We assume that the reviews are generated according to a Bernoulli document model [5.16] that computes the posterior probability of each class based on the distribution of words in a review. In Bernoulli document model, presence and absence of words in a review is consider as a binary vector that presents a point in a space of words. If we have a vocabulary V containing a set of |V| words, then the kth dimension of a reviews vector corresponds to word w_k in the vocabulary. If b_j be the feature vector for the jth sentence S^j, in the review, then the kth element of b_j termed as b_{jk} is either 0 or 1 representing the absence and presence of word w_k in jth sentence of a review. Let $P(w_k|C)$ be the probability of word w_k occurring in a review of class C. The probability of w_k not occurring in a review of class C is given by $(1 - P(w_k|C)$. To classify each unlabeled sentence S^j in a review, we estimate the posterior probability for each class a follows

$$P\left(C\middle|S^j\right) = P\left(C \mid b_j\right)$$

$$P\left(C\middle|S^j\right) = P\left(b_j\middle|C\right) P(C) \tag{5.1}$$

$$P\left(C\middle|S^j\right) = P(C) \prod_{k=1}^{|V|} b_{jk} \left[P(w_k|C) + \left(1 - b_{jk}\right)(1 - P(w_k|C))\right].$$

$$P = \begin{cases} P(w_k|\, C) & \text{if } b_{jk} = 1 \\ \text{otherwise} & \\ (1 - P(w_k|\, C)) & \text{if } b_{jk} = 0 \end{cases} \tag{5.2}$$

5.4.3.2. SVM Model for Sentiment Classification

Support Vector Machine (SVM) is very popular machine learning techniques for the text classification [5.17]. SVM finds an optimal hyperplan represented by vector $\vec{v}$ that separate a review in positive class from a review in negative class. As presented in a Figure 20, a set of

positive and negative labeled review vector $\vec{r}$ is considered to be linear separable if there exists a vector $\vec{v}$ and a scalar b such that the following inequalities are applicable

$$Polarity\ (s) = \begin{cases} Positive & \vec{v} \cdot \vec{r} + b \geq 1 \\ Negative & \vec{v} \cdot \vec{r} + b \leq -1 \\ Neutral & \vec{v} \cdot \vec{r} + b = 0 \end{cases} \quad (5.3)$$

where s is the sentence in single review. The margin of the decision boundary is presented by the distance between the two hyperplan $\vec{v} \cdot \vec{r} + b = 1$ and $\vec{v} \cdot \vec{r} + b = -1$ [ ]. For detail description about SVM, readers are encouraged to study [5.17].

5.4.3.3. Aggregated Semantic Score

After classification phase, each sentence has a sentiment score. Positive sentence are scored within [1, 0], negative sentence are scored within [-1, 0], and neutral sentence are scored as 0. In aggregated semantic score phase, we aggregate the sentiment score of each sentence to obtain the overall score of the entire review. Let R be a set of n reviews $\{r_1, r_2, r_3, \ldots, r_n\}$, and S be a set of n sentences in each reviews $\{s_1, s_2, s_3, \ldots, s_n\}$. The sentiment score calculated for each sentence in a set of reviews as follows

$$Score = \sum_{i=1}^{n} \sum_{j=1}^{m} Sco_{r_i s_j}\ . \quad (5.4)$$

where Sco is the sentiment score calculated for each sentence in a review. Polarity of a review can be defined as

$$Polarity\ (r) = \begin{cases} Positive & if \sum_{i=1}^{n} Sco_{s_i} > 0 \\ Negative & if \sum_{i=1}^{n} Sco_{s_i} < 0 \\ Neutral & if \sum_{i=1}^{n} Sco_{s_i} = 0 \end{cases} \quad (5.5)$$

The neutral sentence can be considers as objective sentences obtaining no information about the venue and can be extracted from the dataset.

5.4.4. Recommendation

In this subsection we discuss in detail the proposed CF-based heuristic recommendation system. In terms of functionality CF-based heuristic recommendation module has three main modules: (a) popularity ranking of reviewers and venues, (b) similarity graph generation among popular reviewers, and (c) recommendation module is responsible for generating the recommendation for a reviewer. The detail functionality of the above mentioned modules is discussed in the following subsections.

5.4.4.1. Reviewer-venue Popularity Ranking

This subsection presents the process of assigning popularity ranking to reviewers and venues. The higher ranked venues and reviewers are known as popular venues and expert reviewers, respectively. HA inference model [4.6] is utilized to perform the ranking for producing a set of experienced reviewers and popular venues. To compute the expert reviewers' and popular venues' scores, the popularity ranking method will generates reviewers-to-venue check-in matrix denoted by M_r. Let $[p_v]$ and $[e_u]$ represent the score matrices for a popular venue and an expert reviewer, respectively. authority score matrices, respectively. The following formulas compute the score for popular venues and expert reviewers [4.10].

$$p_v = M_r^T \times e_u. \tag{5.6}$$

$$e_u = M_r \times p_v. \tag{5.7}$$

If we use $p_v^{<n>}$ and $e_u^{<n>}$ to represent the score of popular venue and expert reviewers at nth iteration, then the following equations generate the score of popular venues and expert reviewers iteratively.

$$p_v^{<n>} = (M_r^T \times M_r) \times p_v^{<n-1>}. \quad (5.8)$$

$$e_u^{<n>} = (M_r \times M_r^T) \times e_u^{<n-1>}. \quad (5.9)$$

The purpose behind using HA method is to generate a subset of reviewers, who have commented popular venues, and a subset of venues that are frequently commented by expert reviewers.

5.4.4.2. Reviewer-venue Similarity Graph Creation

This phase creates similarity graphs among experienced reviewers. The idea is to generate a network of like-minded people (reviewers) who share the similar comments by assigning same semantic score for various venues. The graphs constructed in current phase will be made available for CF-based heuristic recommendation process that utilizes a variant of CF Approach to find an optimal path on the graph. Such a path carries a collective opinion about venues by experienced reviewers who are also most similar to an active reviewer.

The similarity computation between two reviewers in the similarity graph is performed by applying the Pearson Correlation Coefficient (PCC)[1]. The value of PCC ranges between -1 and +1. Positive values indicate that the similarity exists between two reviewers, with highest similarity at 1, whereas negative PCC values means the choices of the two reviewers does not match. PCC is computed by using the following formula.

$$sim(x,y) = \frac{\sum_{v \epsilon S_{xy}} (r_{xv} - \overline{r_l})(r_{yv} - \overline{r_y})}{\sqrt{\sum_{v \epsilon S_{xy}} (r_{xv} - \overline{r_u})^2 \sum_{v \epsilon S_{xy}} (r_{yv} - \overline{r_y})^2}}, \tag{5.10}$$

where

$$S_{ij} = \{v \epsilon V | r_{xv} \neq 0 \wedge r_{yv} \neq 0\}.$$

In (5.10), the similarity between two reviewers x and y is computed only for venues that are commented by both of the reviewers.

The similarity computation in (5.10) results into a very sparse similarity graph due to the fact that majority of the venues are not commented by either of the two reviewers. Therefore, to address the data sparseness problem, we augment the similarity computation with the concern measure. The concern measure can be interpreted as a conditional probability that a venue commented by a one reviewer is also commented by the other reviewer in the dataset. Moreover, it depicts the amount of concern (or confidence) showed by both reviewers in venues commonly commented by them. The following equation is utilized to calculate the weight of an edge between two reviewers.

$$\omega_{ij} = \begin{cases} sim(x,y) & \text{if } sim(x,y) > 0 \\ & \text{otherwise} \\ P(r_x | r_y) \times \frac{1}{1 + \sum_{x \in V_y} |r_{xv} - r_{yv}|} & P[r_y] \neq 0, \end{cases} \tag{5.11}$$

Where V_y is the set of venues checked-in by user y. The parameter $P(r_x | r_y) = P[r_x \cap r_y]/P[r_y]$ is the likelihood ratio that both the reviewers have commented the similar set of venues. The additional sum factor in denominator is used to keep value of probability lower than similarity so that the preference must be given to the positive values of similarity. Moreover, in (5.11), if the similarity value is greater than 0, then this value is assign as an edge weight of

the similarity graph. However, when the similarity value is less than zero, then we consider the lower term of (5.11) to assign the edge weight. This implies that an edge is always assigned a non-zero weight that results in the reduction of data sparseness.

5.4.4.3. Heuristic Recommendation Approach

In this subsection, a heuristic approach is presented that generates a set of top-N venue recommendations based on a graph of the experienced reviewers. The graph of experienced reviewer under a specific feature will be retrieved from the database. The similarity of the active reviewers will be computed with all of the reviewer nodes in the graph using (5.9). Breadth First Search (BFS) procedure will be applied to assign the immediate neighbors of the active reviewer at a distance of one (L = 1). Similarly, neighbors of the active user will be assigned a distance of two ($\boldsymbol{L = 2}$). The process continues until the entire graph is traversed. Each edge of the graph has a weight that is calculated by utilizing cumulative similarity formula described in (5.9). Whereas, the edges connecting the nodes at the same distance are intentionally labeled blank as resented in in Figure 18, because they are not traversed during the execution of Algorithm 3. The top-N venues recommended by the heuristic approach are the one that are not previously visited by the active reviewer. Algorithm 2 illustrates the step-by step procedure of the heuristic approach for online recommendations.

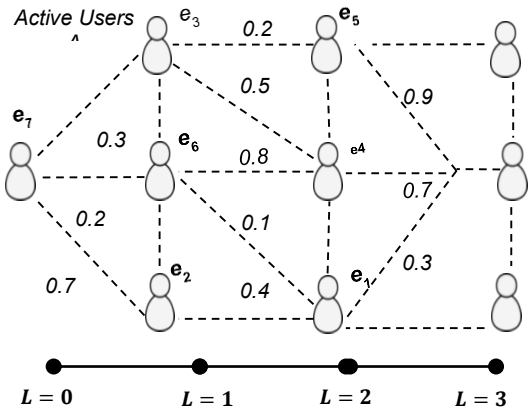

Figure 20. Active users' Similarity graph

a. Initializations (Line 1–Line 5): The identification of the active reviewer, type of venues to be recommended for active reviewer and frequent features of the venue for which reviewer need recommendation are taken as the input of the Algorithm 2.

   In the Line 2 and Line 3, the similarity graph of the experienced reviewers is retrieved. Only those neighbors of active reviewers are selected form the graph that have non-zero similarity computation with the active reviewer. In Line 4, the current reviewer node is stored in the list known as V.

b. Iterative solution construction (Line 5–Line 22): In the Line 5, the weights are assigned to each neighbor nodes ($\mathbf{N_a}$) based on the similarity function $\mathbf{sim}\ (\mathbf{a}, \mathbf{j})$ (defined in (5.9)) that is further multiplied by the $1/\mathbf{\Delta_{rj}}$ that is the edge count between the active reviewer and neighboring node.

   Only those venues are selected from the neighboring nodes that were not previously visited by the active reviewer (Line 7).The selected venues are appended in the matrix A. The visited neighbor is stored in the list V (Line 6–Line 10).

If at Line 11, the venue count in the matrix A is greater than the required number of venues N, then the control jumps to Line 22 that generates the ranking of the venues in the matrix A.

Algorithm 5. Heuristic approach for venue recommendation

Input: Active reviewer : r, Feature : f

Output: A set S' of top-N venues visited by experienced reviewer similar to active reviewer.

Definitions N_j= *neighbor set of node j,* Δ_{ij} = *edge count between reviewers i and j, and,* w_{aj} = *edge weight between reviewer a and j* $\mathcal{Z}_j$= *number of required venues found at a node j, V= list of reviewers visited by active reviewer* r.

1: $a \leftarrow r;\ L \leftarrow 1;\ V \leftarrow \emptyset$
2: $G_f \leftarrow SimGraph(f)$
3: $N_a \leftarrow \{x: G_f | sim(a,x) > 0\}$
4: $V \leftarrow a$
5: $\forall j \in N_a, w_{aj} \leftarrow [sim\ (a,j) \times 1/\Delta_{rj}], j \in N_a$
6: for each $e \in N_a$ do
7: $S \leftarrow \{v: V_e | v \notin V_r\}$
8: $A \leftarrow A.append(e,S)$
9: $V \leftarrow V \cup \{e\}$
10: end for
11: if $venueCount(A) \geq N$ then
12: go to Line 23
13: else
14: $\forall j \in N_a$, select $a \leftarrow j$, such that we have
$arg\,max \left[w_{aj} \times \frac{\mathcal{Z}_j}{N}\right] \wedge\ N_j \neq \emptyset \wedge \forall g \in N_j |\ g \notin V$
15: if $No\ any\ such\ node\ found\ in\ Step$ 14 then
16: go to Line 23
17: else
18: $L \leftarrow L + 1;$
19: go to Line 5
20: end if
21: end if
22: $S' = generaterank(A)$
23: return S'

If the required venue count is not achieved, then new active node (a) is selected amongst the neighbor set N_a. The criterion for the new active nodes selection is that the nodes must have the maximum of the required number of venues. If no such node is found, then the control parses the Line 23. Otherwise, the edge count will also be incremented in Line 18 and in Line 19 and the control will jump back to Line 5.

c. Aggregate venues provided by the best nodes (Line 22): The venues are ranked to generate top-N venues to be recommended to the active user. The following equation is used to rank the venues.

$$Rank_x = \frac{\sum_{e \in V} w(r,e) \times r_{ex}}{\sum_{e \in V} w(r,e)}. \quad (5.12)$$

In (11), x is the venue to be ranked, the parameter r is the active reviewer node, and s_{ex} is the review score calculated for expert reviewer e ∈ V at venue x. The parameter $w(r,e)$ represents the weight of the link in the similarity graph between the root node r and the expert reviewer e.

5.5. Performance Evaluation

In this section, we present the performance evaluation of the proposed BORF. We compare our results with following related schemes: (a) User-Based Collaborative Filtering [5.18] (UBCF), (b) Matrix Factorization (MF) [5.17], and (c) Random Walk with Restart (RWR) [5.6]. A brief description of the schemes is presented in the next subsection.

5.5.1. Results

We utilized “Yelp” dataset that consists of 335023 reviews performed by 150,734 users in total number of 1,280,969 venues [5.6]. Reviewers with least number of visits have been filtered out by setting the threshold on the number of reviewers. The reason for such filtration is that

many of users in “Yelp” dataset provided reviewers at very few places that may not add significant contribution into the real-time analysis. We perform extensive experiments on our internal OpenNebula cloud setup running on 96 core Supermicro SuperServer SYS-7047GR-TRF systems. In the selected dataset, out of the entire records, 80% of the record is used as the training set and 20% constitute the test set for the evaluation. We used a standard 5-fold cross validation technique for evaluating the accuracy rate of the framework [5.3].

We utilized the three popular performance evaluation metrics to evaluate the proposed recommendation frameworks: (a) precision, (b) recall, and (c) F-measure. The precision presents a ratio of the accurate recommendations (true positive (tp)) to the total number of anticipated recommendations (tp+ false positive (fp)). An accurate recommendation is the recommendation

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}. \tag{5.13}$$

that has been predicted correctly in the top-N recommended venues.

The recall measures the completeness by computing the average quality of the individual recommendations. The recall presents the proportion of all the accurate recommendations in the top-N recommended venues and can be represented as:

$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}. \tag{5.14}$$

The F-measure is the harmonic mean of precision and recall and denoted as follows:

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5.15}$$

As presented in the Figure 21, the SocialRec framework achieves the best performance as compared to the rest of the schemes, such as UCF, SVD, RWR, and Popular. Each of the plots presented in the graphs show the average of 150 random runs. The reason of improved

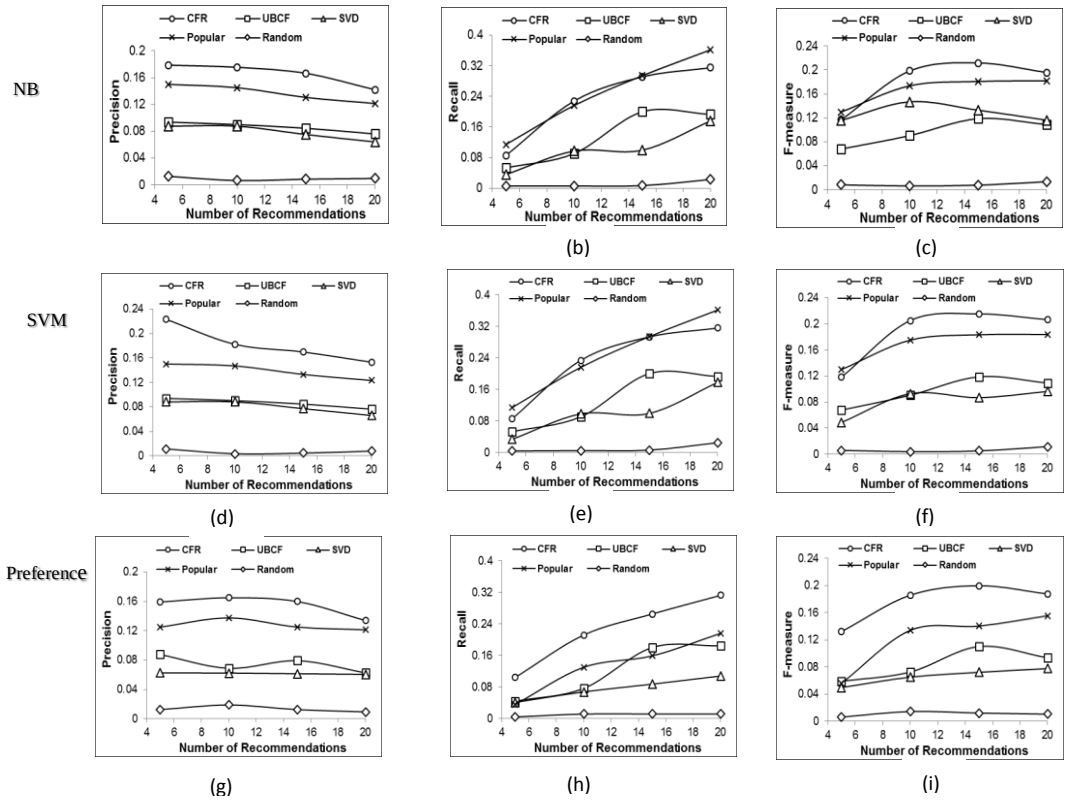

Figure 21. Performance evaluation results: NB (a) Precision, (b) Recall, (c) F-measure : SVM (d) Precision, (e) Recall, and (f) F-measure: Preference (g) Precision, (h) Recall, (i) F measure

performance of SocialRec framework is that the semantic-based recommendation provides more effective solutions towards the data sparsity by taking into account the textual reviews as an addition source of information. Moreover, the SocialRec framework provides improved solution of data sparseness problem by augmenting the similarity computation with conditional probability. Data sparseness results in zero similarity values. The large number of zero entries in user-to-user similarity matrix decreases the recommendation quality. Despite the fact that any two persons have visited almost the same set of venues, the similarity value of the two persons will be smaller or zero if they have significant difference in visit patterns. To reduce the number of zero entries in user-to-user weighted matrix in the aforementioned scenario, we augmented

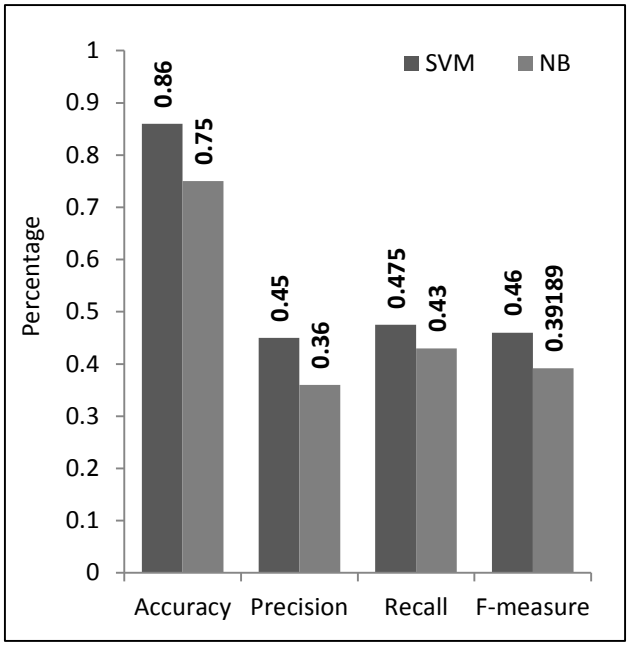

Figure 22. Comparisons between SVM and NB

similarity values with confidence. Therefore, if similarity of two persons is zero but they have visited almost similar set of venues (with different patterns), then they will not be assigned a zero weight in the user-to-user matrix. The reduction in data sparseness results in an increased recommendation precision. The well-known collaborative filtering techniques, such as SVD and UCF presented low performance in terms of precision, recall, and f-measure due to high data sparseness. The popularity-based approach presents comparatively better performance than the collaborative filtering approach. The reason of the improved performance is that the popularity-based approach does not compute the similarity matrix. Therefore, the popularity-based approach is not significantly affected by data sparseness problem. As presented in Figure 21(c), the recall of SocialRec framework is the highest for N=20, which indicates that the framework provides a greater coverage in terms of recommendations. The SocialRec framework indicates better performance in terms of the F-measure as compared to other schemes, such as SVD, UCF,

popular, and random due to the higher values of precision and recall at N=10. The performance of RWR remains low for all the aforementioned metrics.

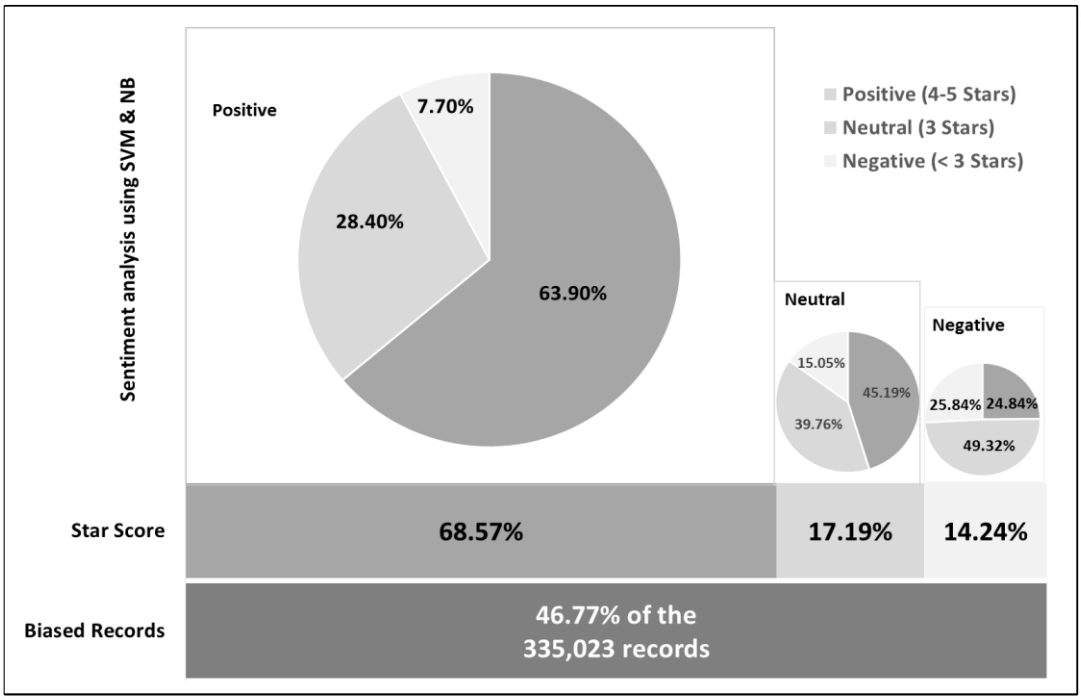

Figure 23. Statistical analysis of positive and negative reviews

Figure 22 presented the comparative analysis of two classification techniques SVM, and naive Bayesian approach. As reflected from the figure SVM achieve better result in terms of accuracy, precision, recall, and f-measure. The Figure 23 presented a statistical analysis of the "Yelp" dataset. Out of 335023 numbers of reviews, there are only 53.23% of reviewers were one that has similar preference rating and sentiment score. In the original dataset there were 68.57% preference score of 4 or 5, 17.9% preference score of 3, and 14.24% preference score of less than 3. The Figure 23 presented the biasness in preference rating of the reviewers. The figure shows that out of 68.57% of positive reviews only 63.9% reviewers were those that are actually positive reviews the rest of 28.4% were identified as neutral reviewers marked as positives reviewers. Similarly, out of 68.57% of positive reviewers 7.7% reviewers were actually negative marked as

positive reviews during preference rating. The figure also depicts that out of 17.19% of neutral reviews indicated by preference rating were 45.19% positive, 39.79% neutral, and 15.05% negative, when evaluated by sentiment score.

5.6. Conclusions

We have presented a multifold contribution by contriving a SocialRec framework that produces recommendations by considering textual data and utilizing the free-text user review for recommendation.

The significance of the proposed framework is the adaptation of popular collaborative filtering and sentiment classification, such as SVM and NB method to compute positive and negative reviews for the recommendation. In our proposed approach, data sparseness issue is address by integrating the user-to-user similarity computation with confidence measure that quantifies the amount of similar interest indicated by the two users in the venues commonly visited by both of them. Moreover, cold start issue is resolves by introducing the HA inference model that assigns ranking to the users and venues based on a mutual reinforcement relationship. In such case, the SocialRec framework always has a precompiled set of popular unvisited venues that can be recommended to the new user.

In the future, we would like to extend our work by incorporating more contextual information, such as the check-in time, users' profiles, and interests, in our proposed framework. Moreover, we intend to integrate other approaches, such as machine learning, text mining, and artificial neural networks to refine our existing framework.

5.7. References

[5.1] J. Bobadilla, F. Ortega, A. Hernando, A. Gutiérrez, "Recommender systems survey," Knowledge-Based Systems, vol. 46, pp. 109-132, July 2013.

[5.2] A. Majid, L. Chen, G. Chen, H. Turab, I. Hussain, and J. Woodward, "A Context-aware Personalized Travel Recommendation System based on Geo-tagged Social Media Data Mining," International Journal of Geographical Information Science, pp. 662-684, 2013.

[5.3] J. J. Ching, E. H. Chan, B. Shi, and V. Tseng, "TripCloud: An Intelligent Cloud-based Trip Recommendation System,"In Proceedings of International Conference on Advances in Spatial and Temporal Databases, pp. 234-256, 2013.

[5.4] Q. Qi, Z. Chen, J. Liu, C. Hui, and Q. Wu, "Using Inferred tag rating to Improve User-based Collaborative Filtering," In Proceedings of the 27th ACM Symposium on Applied Computing, ACM, pp. 2008-2013, 2012.

[5.5] C. Chow, J. BAO, and M. Mokbel, "Towards Location-Based Social Networking Services," In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, ACM, pp. 31-38, 2010.

[5.6] B. Chandra, S. Bhaskar, "Patterned Growth Algorithm using Hub-Averaging without Pre-assigned Weights," In Proceeding of IEEE International Conference on Systems, man, and Cybernetics (SMC), pp.3518-3523, 2010.

[5.7] G. Ganu, Y. Kakodkar, A. Marian, " Improving the Quality of Prediction using Textual Information in Online Review," Information System, vol. 38, no. 1, pp. 1-15, 2013.

[5.8] M. Roy, " Empirical Study of Different classifiers for Sentiment Analysis," Data Mining and Knowledge Engineering, vol. 6, no. 4, 2014.

[5.9] D. Munkova, M. Munk, M. Vozar, “Data Pre-Processing Evaluation for Text Mining: Transaction/ Sequence Model,” Procedia Computer Science, vol. 18, pp. 1198-1207, 2013.

[5.10] O. Khalid, M. U. S. Khan, S. U. Khan, and A. y. Zomaya, “OmniSuggest: A Ubiquitous Cloud based Context aware Recommendation system for Mobile Social Networks,” IEEE Transaction on Services Computing, 2013.

[5.11] J. Atwan, M. Mohd, G. Kanaan, “Enhanced Arabic Information Reterieval: Light Stemming and Stop Words,” Communication in Computer and Information Science, vol. 378, pp. 219-228, 2013.

[5.12] http://pythonhosted.org/pyenchant/

[5.13] R. Forsati, M. Shamsfard, “ Hybrid PoS-Tagging: A cooperation of Evolutionary and Statistical Approach”, Applied Mathematical Modelling, vol. 38, no. 13, pp. 3193-3211, 2014.

[5.14] J. Chung, C. Wu, R. Tzong, “Polarity Detection of Online Reviews Using Sentiment Concept: NCU Team at ESWC-14 Challenge on Concept Level Sentiment Analysis,” Communication in Computer and Information Science, vol. 475, pp. 55-58, 2014.

[5.15] V. Narayanan, I. Aora, A. Bhatia, “Fast and Accurate Sentiment Classification Using Enhanced Naïve Bayes Model,” Intelligent data Engineering and Automated Learning, vol. 8206, pp. 194-201, 2013.

[5.16] P. Pitchandi, N. Raju, “ Improving the Performance of Multivariate Bernoulli Model Based Document Clustering Algorithm using Transformation Technique,” Journal of Computer Science, vol. 7, no. 5, pp. 762-769, 2012.

[5.17] S. Maldonado, G. LHuillier, " SVM-Based Feature Selection and Classification for Email Filtering," Advances in Intelligent Systems and Computing, vol. 204, pp. 135-148, 2013.

[5.18] H. Liu, J. He, T. Wang, W. Song, X. Du, "Combining User Preferences and User Opinion for Accurate Recommendation," Electronic Commerce research and Applications, vol. 12, pp. 14-23, 2013.

[5.19] E. Taylor, J. Velasquez, F. Marquez, Y. Matsuo, "Identifying Customer Preferences about Tourism Products using an Aspect-Based Opinion Mining Approach," In 17th International Conference in Knowledge based and Intelligent Information and Engineering Systems, pp. 182-191, 2013.

[5.20] G. Chen, L. Chen, "Recommendation based on Contextual Opinion," User Modeling, Adaption, and Personalization, vol. 8538, pp. 61-73, 2014.

[5.21] X. Liu, Y. Liu, K. Abeer, " Personalize Point-of-Intrest recommendation by Mining Users' Preference Transition," in 13th proceedings of the 22nd ACM International Conference on Information and Knowledge management, pp. 733-738, 2013.

[5.22] H. P. Hsieh, C. Te, S. Lin, "Exploiting Large-Scale Check-in Data to Recommend Time Sensitive Route," In Proceedings of 12th International Conference of Urban Computing, pp. 55-62, 2012.

[5.23] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "A Random Walk around the City: New Venue Recommendation in Location-Based Social Networks," In Proceedings of International Conference on Social Computing (SocialCom), IEEE, pp.144-153, 2012.

6. CONCLUSION AND FUTURE WORK

This dissertation contributes to the development of novel context-aware techniques to enhance content-based, media-based, and geo-location-based SNSs. Existing social computing systems have emerged through a series of evolutionary steps from single systems to network systems and from network systems to social network systems. The SNSs are attracting researchers' attention due to the significant increase of virtual social interaction. Despite all of the advancements in the content-based, media-based, and geo-location-based SNSs surveyed in this study, SNSs still require further improvements. Context-aware technologies and the increasing diffusion of semantic-based applications offer a new direction to improve present SNSs. This dissertation provides an exclusive overview of the main features and implication of different content-based, media-based, geo-location-based, and context-based SNSs in the various popular SNPs, such as Facebook, Flicker, and LinkedIn. It can be observed that the latest trend is to include the context-based technique into existing SNSs for better, real time, and on-demand communication. The presented ideas will lead the researcher to explore the important research areas, such as on-demand collaboration, on-demand communication, social search, and context-aware recommendation systems.

6.1. Summary of Contributions

In Chapter 3, we presented a detailed analysis of ontology learning and its implication in the field of on text mining. The dynamic visualization of text due to the current Read-Write-Web (RWW) provides an enormous and growing source of information. However, extracting required information and sharing it in different application remain a challenging task. The categorization of unstructured text is one of the fundamental data analysis techniques that have been widely studied in the various disciplines for indexing, mining, and managing abundant textual data.

Ontology offers the potential for providing a logical interpretation of textual data that is based on a hierarchical conceptual representation of information. However, one of the major obstacles that prevents ontology from being deployed in large-scale information systems is ontology acquisition, which strongly depends on knowledge engineers and domain experts. Additionally, ontology building is a labor-intensive, handcrafted, and recursive process. Therefore, to address the abovementioned problem, researchers have devised semi-automatic techniques called ontology learning for building ontologies. This survey provides a comprehensive analysis of ontology learning techniques, such as linguistic, statistical, and semantic-based techniques, extensively used in ontology learning. Moreover, the survey provides a detailed review of the ontology learning process. The discussion moves on further to present the ontology-based text mining architecture and highlights various attempts of scientific researchers to successfully incorporate ontologies in the field of text mining. Furthermore, we identify major issues and challenges in the ontology learning process that need to be addressed in future semantic-based text extraction efforts.

Chapter 4, we presented a multifold contribution by devising cloud based solutions for the venue recommendation problem in mobile social networks for a single user. The novelty and significance of this work was the integration of knowledge engineering techniques, Hub-Average (HA) inference model, multi-obective optimization, and collaborative filtering on a cloud infrastructure to generate optimal set of recommendations. Different from the previous works, the proposed MobiContext, framework not only took into account the collective opinions of the experienced users, but also considers the effect of dynamic real-world physical factors, such as a person’s distance from venues, speed, weather conditions, and travel conditions. The MobiContext utilizes multi-objective optimization techniques to generate personalized

recommendations. The scalability issues were addressed by proposing a cloud-based architecture that allocated data and computational load on geographically distributed cloud nodes. To address the issues pertaining to cold start and data sparseness, the BORF performs data preprocessing by using the. Moreover, the Weighted Sum Approach (WSA) is implemented for scalar optimization and an evolutionary algorithm (NSGA-II) is applied for vector optimization to provide optimal suggestions to the users about a venue. The results of comprehensive experiments on a large-scale real dataset confirm the accuracy of the proposed recommendation framework.

Chapter 5 presents a SocialRec, a context-aware recommendation framework that utilizes a rating inference approach to incorporate textual users' review into traditional collaborative filtering methods for personalized recommendations. To address the issues pertaining to cold start and data sparseness the SocialRec utilizes the textual reviews as an additional source of user preference. Moreover, the SocialRec performs preprocessing by using the Hub-Average (HA) inference model. The results of comprehensive experiments on a large-scale real dataset confirm the accuracy of the proposed recommendation framework.

6.2. Future Work

In this section, we highlight some of the research directions that we intend to explore in future. In the future, we would like to extend our work by incorporating more contextual information in the form of objective functions, such as the check-in time, users' profiles, and interests, in our proposed framework. Moreover, we intend to integrate other approaches, such as machine learning, text mining, and artificial neural networks to refine our existing framework.