ICST Transactions on Scalable Information Systems Research Article ICST.ORG

Personalised Information Gathering and Recommender Systems: Techniques and Trends

Xiaohui Tao[†*], Xujuan Zhou[‡], Cher Han Lau[‡] and Yuefeng Li[‡]

[†]Faculty of Sciences, University of Southern Queensland, Australia
[‡]Science and Engineering Faculty, Queensland University of Technology, Australia

Abstract

With the explosive growth of resources available through the Internet, information mismatching and overload have become a severe concern to users. Web users are commonly overwhelmed by huge volume of information and are faced with the challenge of finding the most relevant and reliable information in a timely manner. Personalised information gathering and recommender systems represent state-of-the-art tools for efficient selection of the most relevant and reliable information resources, and the interest in such systems has increased dramatically over the last few years. However, web personalization has not yet been well-exploited; difficulties arise while selecting resources through recommender systems from a technological and social perspective. Aiming to promote high quality research in order to overcome these challenges, this paper provides a comprehensive survey on the recent work and achievements in the areas of personalised web information gathering and recommender systems. The report covers concept-based techniques exploited in personalised information gathering and recommender systems.

Keywords: Personalisation, Information Gathering, Recommender Systems

Received on 30 December 2011; accepted on 30 April 2012; published on 04 February 2013

doi:10.4108/trans.sis.2013.01-03.e4

1. Introduction

Over the last decade, the rapid growth and adoption of World Wide Web have further exacerbated user needs for efficient mechanisms for information and knowledge location, selection and retrieval. Web information covers a wide range of topics and serves a broad spectrum of communities [2]. Web users create Web information and new sources of knowledge at a rapid rate with various Web 2.0 applications such as blogs, social and professional networks, wikis, and many other types of social media. The abundance of information created by users explicitly and proactively contains rich semantic meaning and provides a huge potential to obtain deep knowledge about users. However, the massive User-Generated Content (UGC) in Web 2.0 era has made it increasingly difficult for users to effectively find exactly what they need. How to gather useful and meaningful information from the Web becomes a challenge to all users. This challenging issue is referred by many researchers as Web information gathering [27, 51].

Web information gathering aims to acquire useful and meaningful information for users from the Web. Web information gathering tasks are usually completed by the systems using keyword-based techniques. The keyword-based mechanism searches the Web by finding the documents with the specific terms matched. This mechanism is used by many existing Web search systems, for example, Google and Yahoo!, for their Web information gathering. Han and Chang [37] pointed out that by using keyword-based search, web information gathering systems can access the information quickly; however, the gathered information may possibly contain much useless and meaningless information. This is particularly referred as the fundamental issue in Web information gathering: information mismatching and information overloading [53–55, 57, 136].

Web-based recommender systems are the most notable application of the web personalization. With today's increasing information overload problem on the Web, the area of recommender systems research

*Corresponding author. xtao@usq.edu.au

becomes more challenging than ever before. There remain difficulties that limit the full exploitation of personalization and resource selection through recommendation from both technology perspective and human and social perspective. Recommender systems have also been made by researchers as an important response to information overloading problems, for its ability to provide personalized and meaningful information recommendations by taking into account idiosyncratic user interests and information needs [90]. For example, while standard search engines are very likely to generate the same results to different users entering identical search queries, recommender systems are able to generate results to each user that are personalized and more relevant because they take into account each user's personal interests. The recommender technology has been successfully employed in many applications such as films, music, books. The richness of the online information challenges the current personalization techniques and also provides new possibilities for accurately users profiling. Thus, how to incorporate the new features and practices of Web 2.0 into personalized recommender applications becomes an important and urgent research topic.

Recommendation techniques can be divided into two major classes: content-based filtering (CBF) and collaborative filtering (CF). CBF focuses on the analysis of item content and user profiles are used to filter available objects. Collaborative filtering (CF) focuses on identification of other users with similar tastes, and utilisetheir opinions to recommend items. The user profiles are used to recommend to a user the information that satisfied previous users with a similar profile. In movie recommender application, for instance, a CBF system will typically rely on information such as genre, actors, director, producer etc. and match this against the learned preferences of the user in order to select a set of promising movie recommendations. CBF recommender systems need a technique to represent the features of the items. Feature representation can be created automatically for machine readable items (such as news or papers). However, for some items such as jokes, it is almost impossible to define the right set of describing features and to "objectively" classify them [73]. Collaborative filtering (CF) collects information about a user by asking them to rate items and makes recommendations based on highly rated items by users with similar taste. CF approaches make recommendations based on the ratings of items by a set of users (neighbours) whose rating profiles are most similar to that of the target user [7]. CF algorithms generally compute the overall similarity or correlation between users, and use that as weight when making recommendations. In book recommendation application, for example, the first step of the CF system is to find the "neighbours" of the target user. The "neighbours" refer to other users who have similar tastes in books (rate the same books similarly). In the second step, only the books that are highly rated by the "neighbours" would be recommended.

In contrast with the content-base approaches, CF techniques rely on the availability of user profiles that capture past ratings histories of users [7] and do not require any human intervention for tagging content because item knowledge is not required. Therefore, the CF techniques can be applied to virtually any kind of items: papers, news, web sites, movies, songs, books, jokes, locations of holidays, stocks and promise to scale well to large item bases [73]. Collaborative filtering is the most widely used approach to build online recommender systems. It has been successfully employed in many applications, such as recommending books, CDs, and other products at *Amazon.com*, Movies by *MovieLens* [1]. Some methods combine both content and collaborative filtering approaches to make recommendations [96].

In the past decade, many researchers have aimed at gathering Web information and make recommendations for users with consideration of their personalised interest and preferences. In these works, Web user profiles are widely used for user modelling and personalization [49], because they reflect the interest and preferences of users [102]. User profiles are defined by Li and Zhong [57] as the interesting topics underlying user information needs. They are used in Web information gathering to describe user background knowledge, to capture user information needs, and to gather personalized Web information for users [31, 37, 57, 113]. This survey paper attempts to review the development of the concept-based, personalized Web information gathering techniques. The review notes the issues in Web personalization, focusing on Web user profiles and user information needs in personalized Web information gathering. The reviewed scholar reports that the concept-based models utilizing user background knowledge are capable of gathering useful and meaningful information for Web users. However, the representation and acquisition of user profiles need to be improved for the effectiveness of Web information gathering. This survey contributes to better understanding of existing Web information gathering systems.

The paper is organized as follows. Section 2 reviews the concept-based techniques employed by Web information gathering and recommender systems; Section 3 discusses the personalisation issues in the context of Web information gathering and recommendation, focusing on user profile representation and acquisition. Information gathering and recommender systems in social networks are discussed in Section 4. Section 5 makes the final remarks for the survey.

2. Exploiting Concepts for Web Information Gathering and Recommender Systems

Concept-based techniques that are used in web information gathering are also widely exploited in recommender systems as well. Recommender systems rely on concept-based techniques to access concepts of products. Concept-based information gathering techniques use semantic concepts extracted from documents and queries. Instead of matching keyword features representing the documents and queries, concept-based techniques attempt to compare the semantic concepts of documents to those of given queries. Similarity of documents to queries is determined by the matching level of their semantic concepts. The semantic concept representation and extraction are two typical issues in the concept-based techniques and are discussed in the following sections.

2.1. Semantic Concept Representation

Semantic concepts have various representations. In some models, concepts are represented by controlled lexicons defined in terminological ontologies, thesauruses, or dictionaries. In some models, they are represented by subjects in domain ontologies, library classification systems, or categorizations. Some models uses data mining techniques for concept extraction, semantic concepts are represented by patterns. The three representations given different strengths and weaknesses.

lexicon-based representation defines the concepts in terms and lexicons that are easily understood by users. WordNet [28] and its variations [9, 48] are typical models employing this kind of concept representation. In these models, semantic concepts are represented by the controlled vocabularies defined in terminological ontologies, thesauruses, or dictionaries. Because these are being controlled, they are also easily utilized by the computational systems. However, when extracting terms to represent concepts for information gathering, some noisy terms may also be extracted because of term ambiguity. As a result, information overloading problem may occur in gathering. Moreover, the lexicon-based representation relies largely on the quality of terminological ontologies, thesaurus, or dictionaries for definitions. However, the manual development of controlled lexicons or vocabularies (like WordNet) is usually costly. The automatic development is efficient, however, in sacrificing the quality of definitions and semantic relation specifications. Consequently, the lexicon-based representation of semantic concepts was reported to be able to improve the information gathering performance in some works [48, 68], but to be degrading the performance in other works [116].

Many Web systems rely upon subject-based representation of semantic concepts for information gathering. In this kind of representation, semantic concepts are represented by subjects defined in knowledge bases or taxonomies, including domain ontologies, digital library systems, and online categorizations. Typical information gathering systems exploiting domain ontologies for concept representation include those developed by Lim *et al.* [65], by Navigli [82], and by Velardi *et al.* [115]. Domain ontologies contain expert knowledge: the concepts described and specified in the ontologies are of high quality. However, expert knowledge acquisition is usually costly in ycapitalization and computation. Moreover, as discussed previously, the semantic concepts specified in many domain ontologies are structured only in the subsumption manner of *super-class* and *sub-class*, rather than the more specific *is-a*, *part-of*, and *related-to*, the ones developed by [31, 46] and [136]. Some attempted to describe more specified relations, like [13, 103] for *is-a*, [33, 92] for *part-of*, and [41] for *related-to* relations only. Tao *et al.* [107, 108] made a further progress from these works and portrayed the basic *is-a*, *part-of*, and *related-to* semantic relations in one single computational model for concept representation.

Also used for subject-based concept representation are the library systems, like Dewey Decimal Classification (DDC) used by [46, 118], Library of Congress Classification (LCC) and Library of Congress Subject Headings (LCSH) [107, 108], and the variants of these systems, such as the "China Library Classification Standard" used by [132] and Alexandria Digital Library (ADL) [117]. These library systems represent the natural growth and distribution of human intellectual work that covers the comprehensive and exhaustive topics of world knowledge [15]. In these systems, concepts are represented by subjects that are defined by librarians and linguists manually under a well-controlled process [15]. Concepts are constructed in taxonomic structure, originally designed for information retrieval from libraries. These are beneficial to the information gathering systems. The concepts are linked by semantic relations, such as subsumption like *super-class* and *sub-class* in the DDC and LCC, and *broader*, *used-for*, and *related-to* in the LCSH. However, information gathering systems using library systems for concept representation largely rely on knowledge bases. The limitations of the library systems, for example, focus on the United States more than on other regions by the LCC and LCSH, would be incorporated by the information gathering systems that use them for concept representation.

The online categorizations are also widely relied on by many information gathering systems for concept representation. Typical online categorization used for

concept representation include the Yahoo! categorization used by [31] and *Open Directory Project*[1] used by [16, 86]. In these categorizations, concepts are represented by categorization subjects and organized in a taxonomical structure. However, the nature of categorizations is in the subsumption manner of one containing another (*super-class* and *sub-class*), but not the semantic *is-a*, *part-of*, and *related-to* relations. Thus, the semantic relations associated with the concepts in such representations are not in adequate details and specific levels. These problems weaken the quality of concept representation and thus the reducing performance of information gathering systems.

Another semantic concept representation in Web information gathering systems is pattern-based representation that uses multiple terms (e.g. phrases) to represent a single semantic concept. Phrases contain more content than any one of their containing terms. Research representing concepts by patterns include Li and Zhong [53–55, 57–59], Wu *et al.* [122, 124, 125], Zhou *et al.* [138–140], Dou *et al.* [22], and Ruiz-Casado *et al.* [93]. However, pattern-based semantic concept representation poses some drawbacks. The concepts represented by patterns can have only subsumption specified for relations. Usually, the relations exist between patterns are specified by investigation of their containing terms, like [57, 125, 138]. If more terms are added into a phrase, to make the phrase more specific, the phrase becomes a sub-class concept of any concepts represented by the sub-phrases in it. Consequently, no specific semantic concepts like *is-a* and *part-of* can be specified and thus some semantic information may be missing in pattern-based concept representations. Another problem of pattern-based concept representation is caused by the length of patterns. The concepts can be adequately specific for discriminating one from others only if the patterns representing the concepts are long enough. However, if the patterns are too long, the patterns extracted from Web documents would be of low frequency and thus, cannot support the concept-based information gathering systems substantially [125]. Although the pattern-based concept representation poses such drawbacks, it is still one of the major concept representations in information gathering systems.

The semantic content of text documents has different representations, such as controlled lexicons, categories, or patterns. The lexicon-based representation of documents is easy to be understood by users or computational systems. With such a representation, text documents are represented by a set of terms chosen from controlled vocabularies defined in terminological ontologies, thesauruses, or dictionaries. However, when extracting lexical descriptors, some noisy terms are also extracted along with meaningful, representative terms, due to term ambiguity problem. The development of terminological ontologies, thesauruses, or dictionaries is also costly in finance, time, and usually requires a large amount of human power involvement. As a result, lexicon-based semantic content representation is ineffective and costly.

Categorizations are also widely used to represent document contents [40, 86, 88, 109]. In such a representation, concepts are represented by categories and organized in a tree or graphic structure. The relationships existing between concept nodes in the structure are explored in order to measure the capacity of a concept describing or representing the semantic content of a document. However, the natural relationship in categorizations is subsumption of one containing another (super-class and sub-class), but not the detailed, specific semantic relations (like is-a, part-of, and related-to). Thus, the concept specification needs to improve toward a more detailed and specific level.

Another representation is pattern-based that uses multiple phrases to represent document contents [23, 57, 61, 140]. However, pattern-based semantic annotation suffers from a problem by the length of patterns. Concepts are specific and discriminating only if patterns are substantially long. However, if a pattern is too long, its frequency would be very low in documents. Consequently, such pattern becomes useless because of poor applicability [60]. In addition, because of using text mining for pattern extraction, the quality of patterns is difficult to control. As a result, noisy patterns are extracted as well as useful patterns.

2.2. Semantic Concept Extraction

Text classification is the process of classifying an incoming stream of documents into categories by using the classifiers learned from the training samples [66]. In generally, text classification problem can be a "binary" classification problem If there are exactly two classes or a "multi-class" problem if there are more than two classes and each document falls into exactly one class or a "multi-label categorization" problem if a document may have more than one associated category in a classification scheme. Multi-label and multi-class tasks are often handled by reducing them to k binary classification tasks, one for each category [128]. The works conducted by Tao's and Yang et al. [129] are about multi-label text classification. The former worked on categorizing library catalogue items into multiple subjects and the latter adopted active learning algorithms for multi-label classification.

[1] http://www.dmoz.org

There are different types of text classifier. Fung *et al.* [29] categorized them into two types: *kernel-based classifiers* and *instance-based classifiers*. Typical kernel-based classifier learning approaches include the *Support Vector Machines* (SVMs) [43] and regression models [98]. These approaches may incorrectly classify many negative samples from an unlabelled set into a positive set, thus causing the problem of information overloading in Web information gathering. Typical instance-based classification approaches include the *K*-Nearest Neighbour (*K*-NN) [19] and its variants, which do not rely upon the statistical distribution of training samples. However, instance-based approaches are not capable of extracting highly accurate positive samples from the unlabeled set. Other research works, such as [31, 88], have a different way of categorizing the classifier learning techniques: *document representations based classifiers*, including SVMs and *K*-NN; and *word probabilities based classifiers*, including Naive Bayesian, decision trees [43] and neural networks used by [133]. These classifier learning techniques have different strengths and weaknesses, and should be chosen based upon the problems they are attempting to solve.

Machine learning for text classification can be categorised into three groups: supervised, semi-supervised, and unsupervised. When there is a set of pre-classified documents available to train classifiers, the process is referred to as supervised classification. Sometimes, samples may be inadequate or insufficient, though available. Such a problem is referred to as semi-supervised text classification. Nguyen and Caruana [83] proposed a semi-supervised approach to address the problem and learned classifiers from only partial label samples (the training documents are pre-classified into a set of possible classes with only one correct class). Fung et al. [29] introduced an approach to learn classifiers from only positive and unlabelled samples, without negative ones. The approach first extracts negative samples from unlabelled set and builds classifiers as usual. Supervised and semi-supervised text classification techniques more or less rely on pre-classified samples to learn classifiers. Yang et al. [130] proposed to build classification model for a target class without associated training samples, by analysing the correlating auxiliary classes.

Text classification techniques are widely used in concept-based Web information gathering systems. Gauch *et al.* [31] described how text classification techniques are used for concept-based Web information gathering. Web users submit a topic associated with some specified concepts. The gathering agents then search for the Web documents that are referred to by the concepts. Sebastiani [98] outlined a list of tasks in Web information gathering to which text classification techniques may contribute: automatic indexing for Boolean information retrieval systems, document organization (particularly in personal organization or structuring of a corporate document base), text filtering, word sense disambiguation, and hierarchical categorization of web pages. Also, as specified by Meretakis *et al.*[75], the Web information gathering areas contributed to by text classification may include sorting emails, filtering junk emails, cataloguing news articles, providing relevance feedback, and reorganizing large document collections. Text classification techniques have been utilized by to classify Web documents into the best matching interest categories, based on their referring semantic concepts [69] .

Some limitations and weaknesses of these text classification techniques employed in concept-based Web information gathering exist. Glover *et al.* [34] pointed out that Web information gathering performance substantially relies on the accuracy of predefined categories. If the arbitration of a given category is wrong, the performance is degraded. Another challenging problem, referred to as "cold start", occurs when there is an inadequate number of training samples available to learning classifiers. Also, as pointed out by Han and Chang [37], concept-based Web information gathering systems rely on an assumption that the content of Web documents is adequate to make descriptions for classification. When the assumption fails, using text classification techniques alone becomes unreliable for Web information gathering systems. The solution to this problem is to use high quality semantic concepts, as argued by Han and Chang [37], and to integrate both text classification and Web mining techniques.

Ontologies have been studied and exploited by many works to facilitate text classification. Gabrilovich and Markovitch [30] enhanced text classification by generating features using domain-specific and common-sense knowledge in large ontologies with hundreds of thousands of concepts. Camous *et al.* [11] also introduced domain-independent method that uses the Medical Subject Headings (MeSH) ontology. The method investigates the inter-concept relationships and represents documents by MeSH subjects. Similarly, Camous' work considers the semantic relations exist in concepts. However, their work focuses only on the medical domain. Whereas, Wang and Domeniconi [119] and Hu et al. [40] derived background knowledge from Wikipedia to represent documents and attempted to deal with the sparsity and high dimensionality problems in text classification. Instead of Wikipedia with free-contributed entries, the approach proposed in [] uses the superior LCSH, which is a world knowledge ontology and has been under continuous development for a hundred years by knowledge engineers.

Many works exploited pattern mining techniques to help build classification models. Malik and Kender [71] proposed the "Democratic Classifier", which is a

pattern-based classification algorithm using short patterns. Their democratic classifier relies on the quality of training samples and cannot deal with the "no training set available" problem. Bekkerman and Matan [3] argued that most of the information on documents can be captured in phrases and proposed a text classification method that employs lazy learning from labelled phrases. The phrases in their work are in fact a special form of sequential patterns.

Web content mining is an emerging field of applying knowledge discovery technology to Web data. Web content mining discovers knowledge from the content of Web documents, and attempts to understand the semantics of Web data [49, 57]. Based on various Web data types, Web content mining can be categorised into Web text mining, Web multimedia data mining (e.g. image, audio, video), and Web structure mining [49]. In this paper, Web information is particularly referred to text documents existing on the Web. Thus, the term "Web content mining" here refers to "Web text content mining", the knowledge discovery from the content of Web text documents. Kosala and Blockeel [49] categorized Web content mining techniques into database views and information retrieval views. From the database view, the goal of Web content mining is to model the Web data so that Web information gathering may be performed based on concepts rather than on keywords. From the information retrieval view, the goal is to improve Web information gathering based on either inferred or solicited Web user profiles. Web content mining contributes significantly to Web information gathering in either view.

Many techniques are applied in Web content mining, including pattern mining, association rules mining, text classification and clustering, and data generalization and summarisation [53, 55]. Li and Zhong [53–55, 57], Wu *et al.* [124], and Zhou *et al.* [138–140] represented semantic concepts by maximal patterns, sequential patterns, and closed sequential patterns, and attempted to discover these patterns for semantic concepts extracted from Web documents. Their experiments reported substantial improvements achieved by their proposed models, in comparison with the traditional *Rocchio*, *Dempster-Shafer*, and probabilistic models. Association rules mining extracts meaningful content from Web documents and discovers their underlying knowledge. Existing models using association rules mining include Li and Zhong [52], Li *et al.* [56], and Yang *et al.* [127], who used the granule techniques to discover association rules; Xu and Li [126] and Shaw *et al.* [100], who attempted to discover concise association rules; and Wu *et al.* [123], who discovered positive and negative association rules. Some works, such as Dou *et al.* [22], attempted to integrate multiple Web content mining techniques for concept extraction. These works were claimed capable of extracting concepts from Web documents and improving the performance of Web information gathering. However, as pointed out by Li and Zhong [54, 55], the existing Web content mining techniques have some limitations. The main problem is that these techniques are incapable of specifying the specific semantic relations (e.g. *is-a* and *part-of*) that exist in the concepts. Their concept extraction needs to be improved for more specific semantic relation specification, considering the fact that the current Web is nowadays moving toward the Semantic Web [4].

3. Personalisation in Web Information Gathering and Recommender Systems

Web user profiles are widely used by Web information systems for user modelling and personalization [49]. User profiles reflect the interests of users [102]. In terms of Web information gathering, user profiles are defined by Li and Zhong [57] as the interesting topics underlying user information needs . Hence, user profiles are used in Web information gathering to capture user information needs from the user submitted queries, in order to gather personalized Web information for users [31, 37, 57, 113].

Web user profiles are categorized by Li and Zhong [57] into two types: the *data diagram* and *information diagram* profiles (also called *behaviour-based profiles* and *knowledge-based profiles* by [76]). The data diagram profiles are usually acquired by analyzing a database or a set of transactions [31, 57, 76, 104, 105]. These kinds of user profiles aim to discover interesting registration data and user profile portfolios. The information diagram profiles are usually acquired by using manual techniques; such as questionnaires and interviews [76, 113], or by using information retrieval and machine-learning techniques [31]. They aim to discover interesting topics for Web user information needs.

Personalized recommender systems have ability to provide meaningful information recommendations [90] by taking into account idiosyncratic user interests and information needs. The representation of user information needs is variously referred to as ï£¡user profilesï£¡, or ï£¡topic profilesï£¡. Recommender systems can be divided into two major classes: content-based filtering [78] and collaborative filtering recommender [47]. Content-based filtering focuses on the analysis of item content. The user profiles are used to filter available objects. Collaborative filtering focuses on identification of other users with similar tastes, and the use of their opinions to recommend items. The user profiles are used to recommend to a user information that satisfied previous users with a similar profile. The recommender systems success depend on large extent on the ability of the learned profiles to represent the users actual interests. Learning a personalized user profile is one

of the most challenging tasks in developing the next generation of information filtering and recommender systems [62, 140].

User profiles are widely used in not only Web information gathering [57, 107, 108], but also personalized Web services [37], personalized recommendations [76], and marketing research [137]. User profile representation and construction are very important issues within these research fields. In this section, the methods and techniques for profiles representation and construction will be discussed.

3.1. User Profile Representation

User profiles have various representations. As defined by [102], user profiles are represented by a previously prepared collection of data reflecting user interests. In many approaches including conventional information gathering and information recommendation, this "collection of data" refers to a set of terms (or vector space of terms) that can be directly used to expand the queries submitted by users [18, 76, 113]. For instance, traditional content-based information filtering uses single-vector or multi-vector models that produce one term-weight or more than one term-weight vectors [98] to represent the relevant information of the topic of likely interest for a user. Such profiles are called term-based user profiles.

These term-based user profiles, however, may cause poor interpretation of user interests to the users, as pointed out by [55, 57]. Also, term-based user profiles suffer from the problems introduced by the keyword-match techniques because many terms are usually ambiguous. Attempting to solve this problem, Li and Zhong [57] represented user profiles by patterns. However, pattern-based user profiles also suffer from problems of inadequate semantic relations specification and the dilemma of pattern length and pattern frequency, as discussed previously in Section 2 for pattern-based concept representation. Recently, the two-stage information filtering (i.e., recommender system) and decision making support system have been developed by Zhou et.al. [140] and Li et.al. [62] using both the term-based and pattern-based profiles.

User profiles can also be represented by personalized ontologies. Tao *et al.* [107, 108], Gauch *et al.* [31], Trajkova and Gauch [113], and Sieg *et al.* [104] represented user profiles by a sub-taxonomy of a predefined hierarchy of concepts. The concepts exist in the concepts are associated with weights indicating the user-perceived interests in these concepts. This kind of user profiles describes user interests explicitly. The concepts specified in user profiles have clear definitions and extents. They are thus excellent for inferences performed to capture user information needs. However, clearly specifying user interests in ontologies is a difficult task, especially for their semantic relations, such as *is-a* and *part-of*. In these aforementioned works, only Tao *et al.* [107, 108] could emphasis these semantic relations in user interest specification.

User profiles can also be represented by a training set of documents, as the user profiles in TREC-11 Filtering Track [89] and the model proposed by Tao *et al.* [106] for acquiring user profiles from the Web. User profiles (the training sets) consist of positive documents that contain user interest topics, and negative documents that contain ambiguous or paradoxical topics. This kind of user profiles describes user interests implicitly, and thus have great flexibility to be used with any concept extraction techniques. The drawback is that noise may be extracted from user profiles as well as meaningful and useful concepts. This may cause an information overloading problem in Web information gathering.

3.2. User Profile Construction

When acquiring user profiles, the content, life cycle, and applications need to be considered [97]. Although user interests are approximate and explicit, it was argued by [31, 57, 107, 108] that they can be specified by using ontologies. The life cycle of user profiles refers to the period that the user profiles are valuable for Web information gathering. User profiles can be long-term or short-term. For instance, persistent and ephemeral user profiles were built by Sugiyama *et al.* [105], based on the long term and short term observation of user behaviour. Applications are also important factors requiring consideration in user profile acquisition. These factors considered in user profile acquisition also define the utilization of user profiles for their contributing areas and period.

User profile acquisition techniques can be categorized into three groups: *interviewing*, *non-interviewing*, and *semi-interviewing* techniques. Interviewing user profiles are entirely acquired using manual techniques; such as questionnaires, interviews, and user classified training sets. Trajkova and Gauch [113] argued that user profiles can be acquired explicitly by asking users questions. One typical model using user-interview profiles acquisition techniques is the TREC-11 Filtering Track model [89]. User profiles are represented by training sets in this model, and acquired by users manually. Users read training documents and assign positive or negative judgements to the documents against given topics. Based upon the assumption that users know their interests and preferences exactly, these training documents perfectly reflect users' interests. However, this kind of user profile acquisition mechanism is costly. Web users have to invest a great deal of effort in reading the documents and providing their opinions and judgements. However, it is unlikely that Web users wish to burden themselves with answering questions

or reading many training documents in order to elicit profiles [55, 57].

The non-interviewing techniques do not involve users directly but ascertain user interests instead. Such user profiles are usually acquired by observing and mining knowledge from user activity and behaviour [57, 101, 105, 113]. Typical model is the personalized, ontological user profiles acquired by [108] using a world knowledge base and user local instance repositories. Some other works, like [31, 113] and [104], acquire non-interviewing ontological user profiles by using global categorizations such as Yahoo! categorization and Online Directory Project. The machine-learning techniques are utilized to analyse the user-browsed Web documents, and classification techniques are used to classify the documents into the concepts specified in the global categorization. As a result, user profiles in these models are a sub-taxonomy of the global categorizations. However, because the categorizations used are not well-constructed ontologies, the user profiles acquired in these models cannot describe the specific semantic relations. Instead of classifying interesting documents into the supervised categorizations, Li and Zhong [55, 57] used unsupervised methods to discover interesting patterns from the user-browsed Web documents, and illustrated the patterns to represent user profiles in ontologies. The model developed by [67] acquired user profiles adaptively, based on the content study of user queries and online browsing history. In order to acquire user profiles, Chirita *et al.* [17] and Teevan *et al.* [111] extracted user interests from the collection of user desktop information such as text documents, emails, and cached Web pages. Makris *et al.* [70] comprised user profiles by a ranked local set of categories and then utilized Web pages to personalize search results for a user. These non-interviewing techniques, however, have a common limitation of ineffectiveness. Their user profiles usually contain much noise and uncertainties because of the use of automatic acquiring techniques.

With the aim of reducing user involvement and improving effectiveness, semi-interviewing user profiles are acquired by semi-automated techniques. This kind of user profiles may be deemed as that acquired by the hybrid mechanism of interviewing and non-interviewing techniques. Rather than providing users with documents to read, some approaches annotate the documents first and attempt to seek user feedback for just the annotated concepts. Because annotating documents may generate noisy concepts, global knowledge bases are used by some user profile acquisition approaches. They extract potentially interesting concepts from the knowledge bases and then explicitly ask users for feedback, like the model proposed by [107]. Also, by using a so-called Quickstep topic ontology, Middleton *et al.* [76] acquired user profiles from unobtrusively monitored behaviour and explicit relevance feedback. The limitation of semi-interviewing techniques is that they largely rely upon knowledge bases for user background knowledge specification.

In recommender systems, the construction of accurate profiles is a key task since accurate profiles enable both content-based filtering (to insure recommendations are appropriate) and collaborative filtering (to insure users with similar profiles are indeed similar) [72]. Current existing user profiling for the recommender systems is mainly using user rating data and selected items' content. Usually, hundreds of thousands of users and items are involved in a recommender system, but only a few items are viewed, selected or rated by users. As Sarwar et.al reported in [95], the density of the available ratings in commercial recommender systems is often less than 1%. Moreover, as for new users, they will start with a blank profile without selecting or rating any items at all. These situations are commonly referred to as the data **sparseness** and **cold start** problem [96]. With the increasing use of recommender systems in e-commerce and social networks, maliciously or unfairly influences to the outcomes of recommender systems by creating false user rating data are also intensified. For example: a simple but effective attack to recommender system is to deliberately create a bunch of fake users with pseudo ratings favour or disfavour to some particular products. With the fake information, user profile data becomes unreal and not reliable hence recommender algorithms are impeded by the sparsity, cold start and malicious data problems.

The user profile information can be input explicitly by users or implicitly gathered by software agents that monitor user activity [32]. Explicit acquisition techniques usually require information such as how users rate or select items; whereas implicit acquisition techniques passively observe user behaviours to discover user interests by inferring from user-system interactions. Currently the user profile information for online recommendation is mainly obtained by analysing usage log data such as users' click streams and navigation patterns. Both the explicit and implicit methods have their respective strengths and weaknesses. Explicit techniques are capable of constructing accurate user profile, because information comes directly from the users (e.g., a user rates the relevance of a set of items). However, they may place an increased cognitive burden on users [79]. Implicit acquisition techniques place little or no burden on the users. However, inferences drawn from user interaction are not always valid, as the indicators of user interests are often erratic [45].

In the Web 2.0 era, people engage in a growing variety and number of Web activities on social websites, from buying on commercial sites, to blogging, to tagging, to online dating, to twittering, to post personal pictures. These interactions can serve as a valuable source of the

usersï£¡ implicit feedback. For example, the tags are pieces of light weight textual information but contain very rich and explicit topic information because users proactively provided these tags. They are independent of the content of the items, which makes it possible to achieve content filtering for any items such as video or music files [140]. For example, in recent years, Liang et.al. [63, 64] developed the personalized item recommendation systems by using tag and item information .

4. Information Gathering and Recommender Systems in Social Web

Although the term "social networking" is being used in new ways since the availability of the digital medium, the concepts behind it have been studied for quite a long time. The modern digital medium technology makes sharing contents, collaborating with others, and connecting with each other to create a community faster, easier and more accessible to a wider population than ever before. Social networking is a type of virtual community that has grown tremendously in popularity over the past few years. The network of users is the platform; the community drives the content. Users actively participate in social networks, upload their personal photos, share their bookmarks, write blogs, and annotate and comment on the information provided by others. They create information, build content, and establish online communities. Nowadays, massive quantities of User Generated Content (UGC) on social networks are available.

Unlike the user rating data which is numeric, the UGC comprises various forms of media and creative works such as text, audio, visual, and combined created by users explicitly and pro-actively. Therefore, it contains rich semantic information and provides a huge potential to obtain deep knowledge about users, items, the various relationships among users and items. The UGC has become an important information resource in addition to traditional website materials. From the UGC information, it is possible to acquire users' opinions, perspectives, or tastes towards items or other users. The growing and readily available user-generated content is rising the new opportunity to construct user profiles accurately compared with the existing personalized recommender techniques, as well as to mitigate the cold start and malicious rating problems considerably.

There has been a tremendous increase in user-generated content (UGC) in the past a few years via the technologies of Web 2.0. It is now well recognized that the user-generated content (e.g., product reviews, tags, forum discussions and blogs) contains valuable user opinions that can be exploited for many applications. By exploiting the UGC more effectively via the use of the latest collaborative filtering, data mining techniques, more accurate and sophisticate user profiles can be built. Based on the enhanced user profiles, high quality and reliable recommendations can be generated. Many significant researches have been done to investigate new strategies available in Web 2.0 framework. In this section, we review some new strategies for social recommender systems.

4.1. Using Tag Information for Recommendation

Like other UGC information, tag is becoming an important information source to profile user's topic interests as well as to describe the content or classification of items. A tag is a keyword that is added to a digital object (e.g. a website, picture or video clip) to describe it, but not as part of a formal classification system. Tags are freely chosen keywords and they are a simple yet powerful tool for organizing, searching and exploring the resources. Compared with other traditional implicit user information such as click stream and web log, the tag information has some distinctive advantages. One important advantage is that tags are pieces of light weighted textual information but contain very rich and explicit topic information since they are given by users explicitly and pro-actively. Another important advantage is that it is independent with the content of the items, which makes it possible to do content filtering for any items such as videos, music files etc. Moreover, tagging behaviour forms a three dimensional relationship among users, items and tags such as the additional implied item-tag, user-tag besides the typical implicit user-item relationship.

However, since there is no restriction or boundary on selecting words for tagging items, the tags used by users are free-style and contain a lot of noise such as semantic ambiguity which means that the same tag name has different meaning for different users, tag synonym which means that different tags actually have the same meaning. Another serious situation of tags is that nearly 60% tags are personal tags that are only used by one user [99]. All these disadvantages of tags bring challenges to make use of tags to profile users' topic preferences accurately or describe the topics of the items correctly. Thus, how to solve these problems caused by the free-style vocabulary of tags is a key issue to improve the accuracy of recommendation systems based on tag information.

The work of Tso-shuter *et al.* [114] extended the user-item matrix to user-item-tag matrix to make collaborative filtering item recommendation. However, this work didn't consider the noise of tags. More recently, the noise of tags has become an important research question. In the recent work of [99], a special tag rating function was used to find user's preferences for tags. Along with the tag preferences, the click streams, tag search history of each user were used to

get user's preferences for items through the inferred tags preferences. However, Sen' work needs various kinds of extra information or special function, which makes the work incomparable and gives restrictions to the application of the work. Moreover, it is difficult to determine the influence of tag information when the click streams, search queries were combined together.

Different from Sen's work the approach proposed by Liang *et al.* [63] makes use of the standard item taxonomy or ontology given by experts to represent each user's tag individually to remove the noise of tags. Item taxonomy is a set of controlled vocabulary terms or topics designed to describe or classify items, which is available for various domains. Because item taxonomy is usually designed and developed by experts, reflecting the common views to the description and classification of items, providing not only a standard vocabulary but also a hierarchical structure to represent the relationships among concepts or categories, it can be used to eliminate the inaccuracy caused by the users' free-style vocabulary in social tags.

4.2. Blogs as a Valuable Information Resource

Social media such as *blog*, *Flickr*, *Youtube*, *Facebook*, and *Twitter* has arisen as the major user-generated media platform in recent years. A blog is a simple web page consisting of brief paragraphs of opinion, information, personal diary entries, or links. People express their opinions, ideas, experiences, thoughts, wishes through these free-form writings. A typical blog post can combine text, images, and links to other blogs, web pages, and other media related to its topic. The individual users show their interest in online opinions about products or services. They share their brand experiences and opinions, positive or negative, regarding any product or service. The vendors of these items are increasingly coming to realize that these consumer voices can potentially wield enormous influence in shaping the opinions of other consumers and they are paying more and more attention to these issues [38]. Currently, many sentiment analysis works are focus on product reviews or movie review [121], [134], [85], [8] on blogs, customer review sites, and Web Pages. The opinion mining and sentiment analysis such as customer opinion summarisation [141] and sentiment analysis of user reviews [21] are possibly as augmentations to recommendation systems [110], since it might behoove such a system not to recommend items that receive a lot of negative feedback. The researchers Joshi and Belsare [44] developed a blog mining program called *BlogHarvest*, which searches for, and extracts, a blogger's interests in order to recommend blogs with similar topics. The program uses classification, links, topic similarity clustering and tagging based on opinion mining to provide these features. The program design is based on the knowledge that blogging communities are not formed randomly, but as a result of shared interests. It is also designed to provide a useful search facility to bloggers while generating large amounts of revenue for advertising services and providers.

4.3. Microblogs as Real Time Information Resource

Twitter is one of the micro-blog service providers that achieves great success. Twitter is a micro-blogging service where users send messages (a.k.a., tweets) to a network of associates from a variety of devices. A tweet is a text-based post and only has 140 characters, which is approximately the length of a typical newspaper headline and subhead [77]. The short messages are very easy and convenience to both sender and reader to share things of interest and communicate their thoughts anywhere and anytime in the world. Today, Twitter has gained popularity with over 200 million users and averaging 1600 tweets sent per second. Twitter consists users from different fields including celebrities (@ladygaga, @justinbieber), national leaders (@barackobama, @kevinrudd), news publishers (@cnn, @ap) to general public. Twitter's user base has grown rapidly and the volumes of messages produced by Twitter everyday is vast. According to [131], in April 2010, Twitter had 106 million registered users, 180 million unique visitors everyday. Users often perform search task in microblog (for example, Twitter) to answer their information need [25]. Searching in micro-blog can be different as compared to traditional web search in the following aspects [112]:

1. Users search twitter for information about people and temporal related information.
2. Twitter search is less varying and can be used to monitor content while web search is used to gain knowledge about a topic.
3. Twitter provides more social content and event-related information while web results are more factual and navigational.
4. Language used in Twitter and Web result is very different.

As more and more users post reviews about products and services they use, or express their political and religious views on Twitter, tweets become a valuable source of peoples opinions and sentiments. Tweets data can be efficiently used to infer people's opinions for marketing or social studies. Given its popularity, Twitter is seen as a potential new form of eWOM (electronic word-of-mouth) marketing by the businesses and organizations concerned with reputation management [42]. Twitter has also been

witnessed as the major online platform where news of significant events were broken such as presidential election debate [20], earthquake [94] and the death of Michael Jackson [35]. Twitter is also used as the primary tool in critical situations when communication channel is limited such as the Iranian election 2009 [10] and Mumbai terrorist attack [84]. Micro-blog, *Twitter* and *Facebook* in particular, has become an important tool for users to share various information from personal updates, question answering [26], news [50] to general babbling [6].

One unique characteristic of Twitter is its rapid response to the change of the Twitter sphere. While it cannot be considered as a reliable information source as compared to authoritative media outlets, its ability to gather emerging topics is impressive. This can be achieved by performing trend analysis and topic detection. Naaman *et al.*[80] analyse the characteristics of emerging trends on Twitter and identify two types of trends: exogenous (broadcast events, global news, important days, physical events) and endogenous (memes, retweets, fan activities). The study also presents five key features: content, interaction, participation, time and social, to collect content aggregation statistics for trend analysis. [12] proposed a 5-steps process to model life cycle of term using a novel aging theory based on user authority, calculated using PageRank algorithm. The emerging term selection is based on nutrition (term quality) and energy (term burstiness). Twitter trends can also be mined with data mining technique such as Kohonen's Self-Organising Map (SOM) to visualise users demographic of trending topics, to reveal underlying pattern and characteristics for decision making [14].

An entry point for micro-blog study is to understand the characteristics of Twitter. [50] conducted a study based on 41.7 million users and 106 million tweets to answer whether Twitter is "social network, or a news media?" They extracted 1.47 billion relationship tuples and revealed that users who topped the chart with over 1 million followers are mostly celebrities (e.g. @oprah, @kimkardashian) or mass media (e.g. @cnn, @nytimes). About 77.9% of the relationships are one-way connection with only 22.1% of reciprocal relationship exists. Moreover, 67.6% of users do not even follow any of their followers at all, which shows a very weak social relationship. Interestingly, another study by Weng *et al.*[120] which based on top-1000 Singapore based twitters listed in *Twitterholic.com* shows the opposite from Kwak et al.'s finding. Study performed by Weng *et al.* revealed that 72.4% of users follow more than 80% of their followers and 80.5% of the users have 80% of their friends follow them back. Both studies agree that Twitter is an excellent news alternatives but the social relationship varies among different user groups.

Apart from trending topic identification, topic modelling can also be used to understand tweet content. Latent Dirichlet allocation (LDA) is one of the popular techniques for its performance and flexibility [5]. LDA is a generative topic model that presents underlying topics as a set of infinite mixture. Each document is considered as a probability distribution of topics and their probabilities can then estimated through sampling methods. Alternatively, Ramage *et al.*[87] uses an variation of LDA, namely Labelled LDA (L-LDA) to model tweets for post ranking and user recommendation task. L-LDA allows mixing labelled topics together with latent topics discovered by original LDA. While it is questionable of the performance of bag-of-words (BOW) model such as TF-IDF in micro-blogs retrieval task, this study shows that performance of TF-IDF and L-LDA are almost identical in ranking task and TF-IDF actually outperforms L-LDA in user recommendation task by about 30%. A combination of L-LDA and TF-IDF improves the result of tweets ranking by a slight 3% but significantly boost the performance in user recommendation task by 66%. This shows that TF-IDF can still be an important feature for micro-blogs task.

Conversely, various studies suggested that LDA may not work on Twitter due to the short length of tweets [39, 120]. One way to overcome such problem is to group tweets together to provide more context. Tweets can be grouped by content and terms [39], or by topics [120] or based on users, as an application of author-topic (AT) model [91]. However, studies show that direct application of AT model does not yield significant improvement as compared to simple term-based approach [39, 135]. A Twitter-LDA model proposed by [135], which considers "a single tweet is usually about a single topic" also shows that content aggregation in Twitter performs better compared to author based aggregation. This might be due to less variation in content-based aggregation than author-based.

While information search in micro-blog is important, its research is still budding. Preliminary works includes the understanding the search type [36] , investigation of hash-tag based retrieval [24], Researchers begin to identify ways to deal with the short length of micro-blogs [81] and query expansion to capture more context [74]. TREC 2011 has also created a dedicated micro-blog track to tackle various ad hoc micro-blogs retrieval problems.

Opinion mining in Twitter is different from the opinion mining from the blogs, review sites or other Web pages. Reviews tend to be longer and more verbose than tweets which may only be a few words long and often contain significant spelling errors. Reviews

usually focus on a specific product or entity and contains little irrelevant information. However, tweets tend to be much more diverse in terms of topics with issues ranging from politics and recent news to religion. As the largest, most well-known, and most popular of the micro-blogging sites, Twitter is an ideal sources for spotting the information about societal interest and general people's opinions. However, there has been little prior opinion mining work in the micro-blogging area since Twitter is relative new technology.

5. Conclusions

This paper has discussed the challenges existing in the current Web information gathering and recommender systems, as well as state-of-the-art techniques employed by them to deal with the challenges. In the recent years, much effort has been dedicated by many research groups on effectively accessing web information. Their results have demonstrated that the key to deal with the challenges is moving the current systems towards concept-based and personalised. Moving on this trend, the researchers have made many great achievements, particularly in personalised information gathering and recommender systems. However, many challenging problems still exist in these areas, for example, how to make breakthrough on the current information gathering and recommender systems on social networks.

Some recent study on micro-blog has also been covered in this paper, including the micro-blog background, application, trend analysis, topic detection, opinion mining, and information gathering. While many of the studies performed on micro-blog related problems are still in its early stage, the preliminary results are promising. Many of the problems are studied extensively in traditional information retrieval field, but the technical difficulty and its applicability when applied in micro-blog are still uncertain. This is due to the short and dynamic nature of micro-blog and the challenge in gathering context. Substantial amount of noise always exist in micro-blogs, but the nature of topic-specific for non-spam messages are definitely indicative and expressive for various micro-blog related task.

The future research direction of information gathering and recommender systems on social networking environment is exciting and vibrant. Network analysis and trend detection models that exploit various twitter characteristics, will be beneficial from the large volume of tweets. Data mining techniques (e.g. pattern mining, association rules) can be applied to further improve the retrieval performance. Lastly, high-level semantic feature such as sentiment analysis and entity detection can then be applied in different real-world applications, which will fully unleash the power of social networks as a valuable, wealthy source of information.

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[2] G. Antoniou and F. van Harmelen. *A Semantic Web Primer*. The MIT Press, 2004.

—— BBB ——

[3] R. Bekkerman and M. Gavish. High-precision phrase-based document classification on a modern scale. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 231–239, New York, NY, USA, 2011. ACM.

[4] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic Web. *Scientific American*, 5:29–37, 2001.

[5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[7] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. pages 43–52, 1998.

[8] S. Brody and N. Elhadad. An unsupervised aspect-sentiment model for online reviews. In *HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[9] A. Budanitsky and G. Hirst. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[10] A. Burns and B. Eltham. Twitter free iran: an evaluation of twitterâĂŹs role in public diplomacy and information operations in iranâĂŹs 2009 election crisis. In *Record of the Communications Policy & Research Forum*, pages 298–310, 2009.

[11] F. Camous, S. Blott, and A. F. Smeaton. Ontology-based medline document classification. In *Proceedings of the 1st international conference on Bioinformatics research and development*, BIRD'07, pages 439–452, Berlin, Heidelberg, 2007. Springer-Verlag.

[12] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[13] S. Cederberg and D. Widdows. Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 111–118, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[14] M. Cheong and V. Lee. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In *SWSM '09: Proceeding of the 2nd ACM workshop on Social web search and mining*, pages 1–8, New York, NY, USA, 2009. ACM.

[15] L. M. Chan. *Library of Congress Subject Headings: Principle and Application*. Libraries Unlimited, 2005.

[16] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschütter. Using ODP metadata to personalize search. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM Press, 2005.

[17] P. A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the Web. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 7–14, 2007.

[18] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *Proceedings of the 11th international conference on World Wide Web*, pages 325–332. ACM Press, Honolulu, Hawaii, USA, 2002.

[19] B. V. Dasarathy, editor. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos: IEEE Computer Society Press, 1990.

[20] N. Diakopoulos and D. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1195–1198. ACM, 2010.

[21] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the Conference on Web Search and Web Data Mining (WSDM)*, 2008.

[22] D. Dou, G. Frishkoff, J. Rong, R. Frank, A. Malony, and D. Tucker. Development of neuroelectromagnetic ontologies(NEMO): a framework for mining brainwave ontologies. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 270–279, 2007.

[23] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 581–590, New York, NY, USA, 2007. ACM Press.

[24] M. Efron. Hashtag retrieval in a microblogging environment. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 787–788, New York, NY, USA, 2010. ACM.

[25] M. Efron. Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 62(6):996–1008, 2011.

[26] M. Efron and M. Winget. Questions are content: A taxonomy of questions in a microblogging environment. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10, 2010.

[27] B. Espinasse, S. Fournier, and F. Freitas. Agent and ontology based information gathering on restricted web domains with AGATHE. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 2381–2386, Brazil, 2008.

[28] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. ISBN: 0-262-06197-X. MIT Press, Cambridge, MA, 1998.

[29] G. P. C. Fung, J. X. Yu, H. Lu, and P. S. Yu. Text classification without negative examples revisit. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):6–20, January 2006.

[30] E. Gabrilovich and S. Markovitch. Feature generation for text categorization using world knowledge. In *Proceedings of The Nineteenth International Joint Conference for Artificial Intelligence*, pages 1048–1053, Edinburgh, Scotland, 2005.

[31] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, 1(3-4):219–234, 2003.

[32] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. User profiles for personalized information access. *The Adaptive Web*, Volume 4321/2007, pp54-89, 2007.

[33] R. Girju, A. Badulescu, and D. Moldovan. Automatic discovery of part-whole relations. *Comput. Linguist.*, 32(1):83–135, 2006.

[34] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using Web structure for classifying and describing Web pages. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 562–569, New York, NY, USA, 2002. ACM Press.

[35] D. Goh and C. Lee. An analysis of tweets in response to the death of michael jackson. In *Aslib Proceedings*, volume 63, pages 432–444. Emerald Group Publishing Limited, 2011.

[36] G. Golovchinsky and M. Efron. Making sense of twitter search. 2010.

[37] J. Han and K.-C. Chang. Data mining for Web intelligence. *Computer*, 35(11):64–70, 2002.

[38] T. Hoffman. Online reputation management is hot – but is it ethical? ComputerWorld, 2 2008.

[39] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[40] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou. Exploiting wikipedia as external knowledge for document clustering. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 389–396, New York, NY, USA, 2009. ACM.

[41] D. Inkpen and G. Hirst. Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics*, 32(2):223–262, 2006.

[42] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci.*, 60(11):2169–2188, 2009.

[43] T. Joachims. Text categorization with Support Vector Machines: learning with many relevant features. In *Proceedings of the 10th European conference on machine learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[44] M. Joshi and N. Belsare. Blogharvest: Blog mining and search framework. In *International Conference on*

Management of Data, Delhi, India, , 2006, December 14-16 2006. Computer Society of India.

[45] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[46] J. D. King, Y. Li, X. Tao, and R. Nayak. Mining World Knowledge for Analysis of Search Engine Content. *Web Intelligence and Agent Systems*, 5(3):233–253, 2007.

[47] J. Konstan et al. Grouplens:Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3) (1997) 77ï£¡87.

[48] H. Kornilakis, M. Grigoriadou, K. Papanikolaou, and E. Gouli. Using WordNet to support interactive concept map construction. In *Proceedings. IEEE International Conference on Advanced Learning Technologies, 2004.*, pages 600–604, 2004.

[49] R. Kosala and H. Blockeel. Web mining research: A survey. *ACM SIGKDD Explorations Newsletter*, 2(1):1–15, 2000.

[50] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[51] Y. Li. Information fusion for intelligent agent-based information gathering. In *WI '01: Proceedings of the First Asia-Pacific Conference on Web Intelligence: Research and Development*, pages 433–437, London, UK, 2001. Springer-Verlag.

[52] Y. Li and N. Zhong. Interpretations of association rules by granular computing. In *Proceedings of IEEE International Conference on Data Mining, Melbourne, Florida, USA*, pages 593–596, 2003.

[53] Y. Li and N. Zhong. Ontology-based Web mining model. In *Proceedings of the IEEE/WIC International Conference on Web Intelligence, Canada*, pages 96–103, 2003.

[54] Y. Li and N. Zhong. Capturing evolving patterns for ontology-based web mining. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 256–263, Washington, DC, USA, 2004. IEEE Computer Society.

[55] Y. Li and N. Zhong. Web Mining Model and its Applications for Information Gathering. *Knowledge-Based Systems*, 17:207–217, 2004.

[56] Y. Li, W. Yang, and Y. Xu. Multi-tier granule mining for representations of multidimensional association rules. In *Proceedings of the Sixth IEEE International Conference on Data Mining*, pages 953–958, 2006.

[57] Y. Li and N. Zhong. Mining Ontology for Automatically Acquiring Web User Information Needs. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):554–568, 2006.

[58] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R. Y. Lau. A two-stage text mining model for information filtering. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1023–1032, New York, NY, USA, 2008. ACM.

[59] Y. Li, S.-T. Wu, and X. Tao. Effective pattern taxonomy mining in text documents. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1509–1510, New York, NY, USA, 2008. ACM.

[60] Y. Li, A. Algarni, S.-T. Wu, and Y. Xu. Mining negative relevance feedback for information filtering. In *Proceedings of the IEEE/WIC/ACM international conference on Web Intelligence*, pages 606–613, 2009.

[61] Y. Li, A. Algarni, and N. Zhong. Mining positive and negative patterns for relevance feature discovery. In *Proceedings of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 753–762, 2010.

[62] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R. Y. Lau. A Two-stage Decision Model for Information Filtering. In *Decision Support System*, 52(2012), p706-716.

[63] H. Liang, Y. Xu, Y. Li, R. Nayak, and L. Weng. Personalized recommender systems integrating social tags and item taxonomy. In *Proc. of WI 09*, 2009.

[64] H. Liang, Y. Xu, Y. Li, R. Nayak, and X. Tao Connecting Users and Items with Weighted Tags for Personalized Item Recommendations. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, 2010.

[65] S.-Y. Lim, M.-H. Song, K.-J. Son, and S.-J. Lee. Domain ontology construction based on semantic relation information of terminology. In *30th Annual Conference of the IEEE Industrial Electronics Society*, volume 3, pages 2213–2217 Vol. 3, 2004.

[66] B. Liu, Y. Dai, X. Li, W. Lee, and P. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the Third IEEE International Conference on Data Mining, ICDM2003*, pages 179–186, 2003.

[67] F. Liu, C. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):28–40, 2004.

[68] S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 266–272, New York, NY, USA, 2004. ACM Press.

[69] Z. Ma, G. Pant, and O. R. L. Sheng. Interest-based personalized search. *ACM Transactions on Information Systems (TOIS)*, 25(1):5, 2007.

[70] C. Makris, Y. Panagis, E. Sakkopoulos, and A. Tsakalidis. Category ranking for personalized search. *Data & Knowledge Engineering*, 60(1):109–125, Jan. 2007.

[71] H. H. Malik and J. R. Kender. Classifying high-dimensional text and web data using very short patterns. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 923–928, Washington, DC, USA, 2008. IEEE Computer Society.

[72] B. Marko and S. Yoav. Fab: Content-Based, Collaborative Recommendation. *Communication of the ACM*, 40(3), pp66-72, 1997

[73] P. Massa and B. Bhattacharjee. Using trust in recommender systems: An experimental analysis. In *In Proceedings of iTrust2004 International Conference*, pages 221–235, 2004.

[74] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. *Advances in Information Retrieval*, pages 362–367, 2011.

[75] D. Meretakis, D. Fragoudis, H. Lu, and S. Likothanassis. Scalable association-based text classification. In *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*, pages 5–11, New York, NY, USA, 2000. ACM Press.

[76] S. E. Middleton, N. R. Shadbolt, and D. C. D. Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):54–88, 2004.

[77] S. Milstein, A. Chowdhury, G. Hochmuth, B. Lorica, and R. Magoulas. Twitter and the micro-messaging revolution: Communication, connections, and immediacyÂřx140 characters at a time. An O'Reilly Radar Report . 54 pages, November 2008.

[78] R. Mooney and L. Roy. Information filtering based on user behavior analysis and best match text retrieval. *Proceedings of 5th ACM Conference on Digital Libraries*, pages 195ï£¡204, 2002.

[79] M. Morita and Y. Shinoda. Content-based book recommending using learning for text categorization. *Proceedings of SIGIR '94 ACM*, pages 272–281, 1994.

[80] M. Naaman, H. Becker, and L. Gravano. Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology*, 2011.

[81] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Searching microblogs: coping with sparsity and document quality. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 183–188, New York, NY, USA, 2011. ACM.

[82] R. Navigli, P. Velardi, and A. Gangemi. Ontology learning and its application to automated terminology translation. *Intelligent Systems, IEEE*, 18:22–31, 2003.

[83] N. Nguyen and R. Caruana. Classification with partial labels. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–559, New York, NY, USA, 2008. ACM.

[84] O. Oh, M. Agrawal, and H. Rao. Information control and terrorism: Tracking the mumbai terrorist attack through twitter. *Information Systems Frontiers*, 13:33–43, 2011. 10.1007/s10796-010-9275-8.

[85] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[86] G. Qiu, K. Liu, J. Bu, C. Chen, and Z. Kang. Quantify query ambiguity using odp metadata. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 697–698, New York, NY, USA, 2007. ACM Press.

[87] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2010.

[88] D. Ravindran and S. Gauch. Exploiting hierarchical relationships in conceptual search. In *Proceedings of the 13th ACM international conference on Information and Knowledge Management*, pages 238–239, New York, USA, 2004. ACM Press.

[89] S. E. Robertson and I. Soboroff. The TREC 2002 filtering track report. In *Text REtrieval Conference*, 2002.

[90] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *CSCW*, pages 175–186, 1994.

[91] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[92] D. A. Ross and R. S. Zemel. Learning parts-based representations of data. *The Journal of Machine Learning Research*, 7:2369–2397, 2006.

[93] M. Ruiz-Casado, E. Alfonseca, and P. Castells. Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from Wikipedia. *Data & Knowledge Engineering*, 61(3):484–499, June 2007.

[94] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 851–860, New York, NY, USA, 2010. ACM. p851-sakaki.pdf.

[95] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

[96] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260, New York, NY, USA, 2002. ACM.

[97] J. Schuurmans, B. de Ruyter, and H. van Vliet. User profiling. In *CHI '04: CHI '04 extended abstracts on Human factors in computing systems*, pages 1739–1740, New York, NY, USA, 2004. ACM Press.

[98] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.

[99] S. Sen, J. Vig, and J. Riedl. Tagommenders: Connecting users to items through tags. In *Proc. of WWW' 09*, pages 671–680, 2009.

[100] G. Shaw, Y. Xu, and S. Geva. Deriving non-redundant approximate association rules from hierarchical datasets. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1451–1452, New York, NY, USA, 2008. ACM.

[101] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 824–831, New York, NY, USA, 2005. ACM Press.

[102] M. A. Shepherd, A. Lo, and W. J. Phillips. A study of the relationship between user profiles and user queries. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 274–281, 1985.

[103] K. Shinzato and K. Torisawa. Extracting hyponyms of prespecified hypernyms from itemizations and headings in web documents. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 938, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[104] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 525–534, New York, NY, USA, 2007. ACM.

[105] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web*, pages 675–684, 2004.

[106] X. Tao, Y. Li, N. Zhong, and R. Nayak. Automatic Acquiring Training Sets for Web Information Gathering. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 532–535, HK, China, 2006.

[107] X. Tao, Y. Li, N. Zhong, and R. Nayak. Ontology mining for personalized web information gathering. In *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 351–358, 2007.

[108] X. Tao, Y. Li, N. Zhong, and R. Nayak. An ontology-based framework for knowledge retrieval. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 510–517, 2008.

[109] X. Tao, Y. Li, and N. Zhong. A personalized ontology model for web information gathering. *IEEE Transactions on Knowledge and Data Engineering, IEEE computer Society Digital Library. IEEE Computer Society*, 23(4):496–511, 2011.

[110] J. Tatemura. Virtual reviewers for collaborative exploration of movie reviews. In *Proceedings of Intelligent User Interfaces (IUI)*, pages 272–275, 2000.

[111] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456, 2005.

[112] J. Teevan, D. Ramage, and M. Morris. #twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM, 2011.

[113] J. Trajkova and S. Gauch. Improving ontology-based user profiles. In *Proceedings of RIAO 2004*, pages 380–389, 2004.

[114] Tso-Sutter, K.H.L., L. Marinho, and L. Schmidt-Thieme. Tag-aware recommender systems by fusion of collaborative filtering algorithms. In *Proc. of Applied Computing*, pages 1995–1999, 2008.

[115] P. Velardi, P. Fabriani, and M. Missikoff. Using text processing techniques to automatically enrich a domain ontology. In *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems*, pages 270–284, New York, NY, USA, 2001. ACM Press.

[116] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. Petrakis, and E. E. Milios. Semantic similarity methods in WordNet and their application to information retrieval on the Web. In *WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16, New York, NY, USA, 2005. ACM Press.

[117] J. Wang and N. Ge. Automatic feature thesaurus enrichment: extracting generic terms from digital gazetteer. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 326–333, New York, NY, USA, 2006. ACM.

[118] J. Wang and M. C. Lee. Reconstructing DDC for interactive classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 137–146, New York, NY, USA, 2007. ACM.

[119] P. Wang and C. Domeniconi. Building semantic kernels for text classification using wikipedia. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–721, New York, NY, USA, 2008. ACM.

[120] J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[121] D. T. Wijaya and S. Bressan. A random walk on the red carpet: rating movies with user reviews and pagerank. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 951–960. ACM, 2008.

[122] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and C. P. Automatic pattern taxonomy exatraction for web mining. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, pages 242–248, Beijing, China, 2004.

[123] X. Wu, C. Zhang, and S. Zhang. Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems (TOIS)*, 22(3):381–405, 2004.

[124] S.-T. Wu, Y. Li, and Y. Xu. Deploying approaches for pattern refinement in text mining. In *Proceedings of the Sixth International Conference on Data Mining*, pages 1157–1161, 2006.

[125] S.-T. Wu. *Knowledge Discovery Using Pattern Taxonomy Model in Text Mining*. PhD thesis, Faculty of Information Technology, Queensland University of Technology, 2007.

[126] Y. Xu and Y. Li. Generating concise association rules. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 781–790, New York, NY, USA, 2007. ACM.

[127] W. Yang, Y. Li, J. Wu, and Y. Xu. Granule mining oriented data warehousing model for representations of multidimensional association rules. *International Journal of Intelligent Information and Database Systems*, 2(1):125–145, 2008.

[128] Y. Yang and T. Joachims. Text Categorization. *Scholarpedia*, 3(5):4242, 2008.

[129] B. Yang, J.-T. Sun, T. Wang, and Z. Chen. Effective multi-label active learning for text classification. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 917–926, New York, NY, USA, 2009. ACM.

[130] T. Yang, R. Jin, A. K. Jain, Y. Zhou, and W. Tong. Unsupervised transfer classification: application to text

categorization. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1159–1168, New York, NY, USA, 2010. ACM.

[131] J.Yarow. Twitter finally reveals all its secret stats. BusinessInsider Weblog Article, http://www.businessinsider.com/twitter-stats-2010-4/, 04 2010.

[132] Z. Yu, Z. Zheng, S. Gao, and J. Guo. Personalized information recommendation in digital library domain based on ontology. In *IEEE International Symposium on Communications and Information Technology, 2005. ISCIT 2005.*, volume 2, pages 1249–1252, 2005.

[133] L. Yu, S. Wang, and K. K. Lai. An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering*, 18(2):217–230, 2006.

[134] W. Zhang, C. Yu, and W. Meng. Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 831–840. ACM, 2007.

[135] W. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. *Advances in Information Retrieval*, pages 338–349, 2011.

[136] N. Zhong. Representation and construction of ontologies for Web intelligence. *International Journal of Foundation of Computer Science*, 13(4):555–570, 2002.

[137] N. Zhong. Toward Web Intelligence. In *Proceedings of 1st International Atlantic Web Intelligence Conference*, pages 1–14, 2003.

[138] X. Zhou, Y. Li, P. Bruza, Y. Xu, and R. Y. Lau. Pattern taxonomy mining for information filtering. In *AI '08: Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence*, pages 416–422, Berlin, Heidelberg, 2008. Springer-Verlag.

[139] X. Zhou, Y. Li, P. Bruza, Y. Xu, and R. Y. K. Lau. Two-stage model for information filtering. In *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 685–689, Sydney, Australia, 2008. IEEE Computer Society.

[140] X. Zhou, Y. Li, P. Bruza, Y. Xu, and R. Y. K. Lau. Pattern Mining for a Two-stage Information Filtering System. In *Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2011)*, pages p363-374, Shenzhen, China, 2011.

[141] L. Zhuang, F. Jing, X. Zhu, and L. Zhang. Movie review mining and summarization. In *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*, 2006.