Concept-based, Personalized Web Information Gathering: A Survey

Xiaohui Tao and Yuefeng Li

SIT, Queensland University of Technology, Australia
{x.tao, y2.li}@qut.edu.au

Abstract. Web information gathering surfers from the problems of information mismatching and overloading. In an attempt to solve these fundamental problems, many works have proposed to use concept-based techniques to perform personalized information gathering for Web users. These works have significantly improved the performance of Web information gathering systems. In this paper, a survey is conducted on these works. The reviewed scholar report that the concept-based, personalized techniques can gather more useful and meaningful information for Web users. The survey also suggests that improvement is needed for the representation and acquisition of user profiles in personalized Web information gathering.

1 Introduction

Over the last decade, the rapid growth and adoption of the World Wide Web have further exacerbated user need 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 [1]. How to gather useful and meaningful information from the Web, however, becomes challenging to Web users. This challenging issue is referred by many researchers as Web information gathering [23, 11].

Given an information needs, Web information gathering aims to acquire useful and meaningful information for users from the Web. The 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!information gathering. Han and Chang [17] pointed out that by using keyword-based search techniques, the 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 [27–30, 71].

In attempting to solve these fundamental problems, many researchers have aimed at gathering personalized Web information for users with better effectiveness and efficiency. These researchers have not only moved information gathering

from keyword-based to concept-based, but also take user background knowledge into consideration. In these works, Web user profiles are widely used for user modelling and personalization [22], because they reflect the interest and preferences of users [50]. User profiles are defined by Li and Zhong [30] 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 [14, 17, 30, 58].

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 has contributions to better understanding of existing Web information gathering systems.

The paper is organized as follows. Section 2 reviews the concept-based Web information gathering techniques, including concept representation and extraction. Section 3 presents the survey of personalized Web information gathering, including user profile representation and acquisition. Finally, Section 4 makes the final remarks for the survey.

2 Concept-based Web Information Gathering

The concept-based information gathering techniques use the semantic concepts extracted from documents and queries. Instead of matching the keyword features representing the documents and queries, the concept-based techniques attempt to compare the semantic concepts of documents to those of given queries. The 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, these concepts are represented by controlled lexicons defined in terminological ontologies, thesauruses, or dictionaries. In some other models, they are represented by subjects in domain ontologies, library classification systems, or categorizations. In some models using data mining techniques for concept extraction, semantic concepts are represented by patterns. The three representations given have different strengthes and weaknesses.

The lexicon-based representation defines the concepts in terms and lexicons that are easily understood by users. WordNet [12] and its variations [3, 21] 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 the term ambiguity problem. As a result, the 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 [21, 35], but to be degrading the performance in other works [59].

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 utilizing domain ontologies for concept representation include those developed by Lim *et al.* [32], by Navigli [40], and by Velardi *et al.* [60]. 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 both capitalization 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 or used by [14, 20] and [71]. Some attempted to describe more specified relations, like [4, 51] for *is-a*, [15, 44] for *part-of*, and [18] for *related-to* relations only. Tao *et al.* [55, 56] 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 [20, 62], Library of Congress Classification (LCC) and Library of Congress Subject Headings (LCSH) [55, 56], and the variants of these systems, such as the "China Library Classification Standard" used by [70] and the Alexandria Digital Library (ADL) used by [61]. These library systems represent the natural growth and distribution of human intellectual work that covers the comprehensive and exhaustive topics of world knowledge [5]. In these systems, the concepts are represented by the subjects that are defined by librarians and linguists manually under a well-controlled process [5]. The 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, the information gathering systems using library systems for concept representation largely rely upon the existing knowledge bases. The limitations of the library systems, for example, the 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 upon by many information gathering systems for concept representation. The typical online categorizations used for concept representation include the Yahoo! categorization used by [14] and *Open Directory Project*[1] used by [7, 41]. 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 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 [27–30, 24, 31], Wu *et al.* [65, 64, 63], Zhou *et al.* [73, 74], Dou *et al.* [10], and Ruiz-Casado *et al.* [45]. However, pattern-based semantic concept representation poses some drawbacks. The concepts represented by patterns can have only subsumption specified for relations. Usually, the relations existing between patterns are specified by investigation of their containing terms, like [30, 63, 73]. If more terms are added into a phrase, making 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 missed 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 [63]. Although the pattern-based concept representation poses such drawbacks, it is still one of the major concept representations in information gathering systems.

2.2 Semantic Concept Extraction

The techniques used for concept extraction from text documents include text classification techniques and Web content mining techniques, including asso-

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

ciation rules mining and pattern mining. These techniques are reviewed and discussed as follows.

Text classification is the process of classifying an incoming stream of documents into categories by using the classifiers learned from the training samples [33]. Text classification techniques can be categorized into different groups. Fung *et al.* [13] categorized them into two types: *kernel-based classifiers* and *instance-based classifiers*. Typical kernel-based classifier learning approaches include the *Support Vector Machines* (SVMs) [19] and regression models [47]. These approaches may incorrectly classify many negative samples from an unlabeled 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 Neighbor (K-NN) [9] and its variants, which do not relay upon the statistical distribution of training samples. However, the instance-based approaches are not capable of extracting highly accurate positive samples from the unlabeled set. Other research works, such as [14, 42], 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 [19] and neural networks used by [69]. These classifier learning techniques have different strengthes and weaknesses, and should be chosen based upon the problems they are attempting to solve.

Text classification techniques are widely used in concept-based Web information gathering systems. Gauch *et al.* [14] 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 [47] 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.*[38], 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 [36] to classify Web documents into the best matching interest categories, based on their referring semantic concepts.

Text classification techniques utilized for concept-based Web information gathering, however, incorporate some limitations and weaknesses. Glover *et al.* [16] pointed out that the 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 [17], the concept-based Web information gathering systems rely on an assumption that the content of Web documents is adequate to make descriptions for classi-

fication. When the assumption is not true, 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 [17], and to integrate both text classification and Web mining techniques.

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 [22, 30]. Based on various Web data types, Web content mining can be categorized into Web text mining, Web multimedia data mining (e.g. image, audio, video), and Web structure mining [22]. In this paper, Web information is particularly referred to as the 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 [22] 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. With either view, Web content mining contributes significantly to Web information gathering.

Many techniques are utilized in Web content mining, including pattern mining, association rules mining, text classification and clustering, and data generalization and summarization [27, 29]. Li and Zhong [27–30], Wu *et al.* [64], and Zhou *et al.* [73, 74] 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 [26], Li *et al.* [25], and Yang *et al.* [68], who used the granule techniques to discover association rules; Xu and Li [67] and Shaw *et al.* [48], who attempted to discover concise association rules; and Wu *et al.* [66], who discovered positive and negative association rules. Some works, such as Dou *et al.* [10], 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 [28, 29], the existing Web content mining techniques incorporate 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 [2].

3 Personalized Web information Gathering

Web user profiles are widely used by Web information systems for user modelling and personalization [22]. User profiles reflect the interests of users [50]. In terms of Web information gathering, user profiles are defined by Li and Zhong [30] 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 [14, 17, 30, 58].

Web user profiles are categorized by Li and Zhong [30] into two types: the *data diagram* and *information diagram* profiles (also called *behavior-based profiles* and *knowledge-based profiles* by [39]). The data diagram profiles are usually acquired by analyzing a database or a set of transactions [14, 30, 39, 52, 53]. 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 [39, 58], or by using information retrieval and machine-learning techniques [14]. They aim to discover interesting topics for Web user information needs.

3.1 User Profile Representation

User profiles have various representations. As defined by [50], user profiles are represented by a previously prepared collection of data reflecting user interests. In many approaches, 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 [8, 39, 58]. These term-based user profiles, however, may cause poor interpretation of user interests to the users, as pointed out by [29, 30]. Also, the 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 [30] represented user profiles by patterns. However, the pattern-based user profiles also suffer from the 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.

User profiles can also be represented by personalized ontologies. Tao *et al.* [55, 56], Gauch *et al.* [14], Trajkova and Gauch [58], and Sieg *et al.* [52] represented user profiles by a sub-taxonomy of a predefined hierarchy of concepts. The concepts existing in the taxonomy 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.* [55, 56] 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 [43] and the model proposed by Tao

et al. [54] 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 Acquisition

When acquiring user profiles, the content, life cycle, and applications need to be considered [46]. Although user interests are approximate and explicit, it was argued by [55, 56, 30, 14] 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.* [53], based on the long term and short term observation of user behavior. Applications are also an important factor requiring consideration in user profile acquisition. User profiles are widely used in not only Web information gathering [55, 56, 30], but also personalized Web services [17], personalized recommendations [39], automatic Web sites modifications and organization, and marketing research [72]. 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: the *interviewing*, *non-interviewing*, and *semi-interviewing* techniques. The interviewing user profiles are entirely acquired using manual techniques; such as questionnaires, interviews, and user classified training sets. Trajkova and Gauch [58] 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 [43]. 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 [29, 30].

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 behavior [30, 49, 53, 58]. Typical model is the personalized, ontological user profiles acquired by [56] using a world knowledge base and user local instance repositories. Some other works, like [14, 58] and [52], 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 analyze 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, the 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 [29, 30] 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 [34] 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.* [6] and Teevan *et al.* [57] extracted user interests from the collection of user desktop information such as text documents, emails, and cached Web pages. Makris *et al.* [37] 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 improve effectiveness, the 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 [55]. Also, by using a so-called Quickstep topic ontology, Middleton *et al.* [39] acquired user profiles from unobtrusively monitored behavior and explicit relevance feedback. The limitation of semi-interviewing techniques is that they largely rely upon knowledge bases for user background knowledge specification.

4 Remarks

This survey introduced the challenges existing in the current Web information gathering systems, and described how the current works related to the challenges. The scholar reviewed in this survey suggested that the key to gathering meaningful and useful information for Web users is to improve the Web information gathering techniques from keyword-based to concept-based, and from general to personalized. The concept-based systems using user background knowledge were reported capable of gathering useful and meaningful information for Web users. However, research gaps exist for the representation and acquisition of user profiles, in terms of effective user information need capture.

References

1. G. Antoniou and F. van Harmelen. *A Semantic Web Primer*. The MIT Press, 2004.
2. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic Web. *Scientific American*, 5:29–37, 2001.
3. A. Budanitsky and G. Hirst. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.
4. 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, 2003.
5. L. M. Chan. *Library of Congress Subject Headings: Principle and Application*. Libraries Unlimited, 2005.
6. 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.
7. 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. 2005.
8. 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. 2002.
9. B. V. Dasarathy, editor. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos: IEEE Computer Society Press, 1990.
10. 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.
11. 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, 2008.
12. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.
13. 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.
14. S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, 1(3-4):219–234, 2003.
15. R. Girju, A. Badulescu, and D. Moldovan. Automatic discovery of part-whole relations. *Comput. Linguist.*, 32(1):83–135, 2006.
16. 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, 2002.
17. J. Han and K.-C. Chang. Data mining for Web intelligence. *Computer*, 35(11):64–70, 2002.
18. D. Inkpen and G. Hirst. Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics*, 32(2):223–262, 2006.
19. 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, 1998.
20. 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.
21. H. Kornilakis, M. Grigoriadou, K. Papanikolaou, and E. Gouli. Using WordNet to support interactive concept map construction. In *Proceedings of IEEE International Conference on Advanced Learning Technologies, 2004.*, pages 600–604, 2004.
22. R. Kosala and H. Blockeel. Web mining research: A survey. *ACM SIGKDD Explorations Newsletter*, 2(1):1–15, 2000.
23. 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, 2001.
24. 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, 2008.
25. 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.
26. 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.
27. 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.

28. 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, 2004.
29. Y. Li and N. Zhong. Web Mining Model and its Applications for Information Gathering. *Knowledge-Based Systems*, 17:207–217, 2004.
30. 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.
31. 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, 2008.
32. 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.
33. 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.
34. 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.
35. 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, 2004.
36. Z. Ma, G. Pant, and O. R. L. Sheng. Interest-based personalized search. *ACM Transactions on Information Systems (TOIS)*, 25(1):5, 2007.
37. C. Makris, Y. Panagis, E. Sakkopoulos, and A. Tsakalidis. Category ranking for personalized search. *Data & Knowledge Engineering*, 60(1):109–125, Jan. 2007.
38. 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, 2000.
39. 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.
40. R. Navigli, P. Velardi, and A. Gangemi. Ontology learning and its application to automated terminology translation. *Intelligent Systems, IEEE*, 18:22–31, 2003.
41. 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, 2007.
42. 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, 2004.
43. S. E. Robertson and I. Soboroff. The TREC 2002 filtering track report. In *Text REtrieval Conference*, 2002.
44. D. A. Ross and R. S. Zemel. Learning parts-based representations of data. *The Journal of Machine Learning Research*, 7:2369–2397, 2006.
45. 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.
46. 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, 2004.
47. F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.
48. 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, 2008.
49. 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, 2005.
50. 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.
51. 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.
52. 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, 2007.
53. 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.

54. 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, 2006.
55. X. Tao, Y. Li, N. Zhong, and R. Nayak. Ontology mining for personalzied web information gathering. In *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 351–358, 2007.
56. 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.
57. 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.
58. J. Trajkova and S. Gauch. Improving ontology-based user profiles. In *Proceedings of RIAO 2004*, pages 380–389, 2004.
59. 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, 2005.
60. 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, 2001.
61. 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, 2006.
62. 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, 2007.
63. S.-T. Wu. *Knowledge Discovery Using Pattern Taxonomy Model in Text Mining*. PhD thesis, Faculty of Information Technology, Queensland University of Technology, 2007.
64. 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.
65. 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, 2004.
66. 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.
67. 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, 2007.
68. 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.
69. 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.
70. 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.
71. N. Zhong. Representation and construction of ontologies for Web intelligence. *International Journal of Foundation of Computer Science*, 13(4):555–570, 2002.
72. N. Zhong. Toward Web Intelligence. In *Proceedings of 1st International Atlantic Web Intelligence Conference*, pages 1–14, 2003.
73. 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, 2008.
74. 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, 2008.