Li L, Wang DD, Zhu SZ *et al.* Personalized news recommendation: A review and an experimental investigation. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 26(5): 754–766 Sept. 2011. DOI 10.1007/s11390-011-0175-2

Personalized News Recommendation: A Review and an Experimental Investigation

Lei Li[1] (李 磊), Ding-Ding Wang[1,*] (王丁丁), Shun-Zhi Zhu[2] (朱顺痣), and Tao Li[1,**] (李 涛)

[1] *School of Computing and Information Sciences, Florida International University, Miami, Florida 33199, U.S.A.*

[2] *Department of Computer Science and Technology, Xiamen University of Technology, Xiamen 361024, China*

E-mail: {lli003, dwang003}@cs.fiu.edu; szzhu@xmut.edu.cn; taoli@cs.fiu.edu

Received February 21, 2011; revised June 14, 2011.

Abstract Online news articles, as a new format of press releases, have sprung up on the Internet. With its convenience and recency, more and more people prefer to read news online instead of reading the paper-format press releases. However, a gigantic amount of news events might be released at a rate of hundreds, even thousands per hour. A challenging problem is how to efficiently select specific news articles from a large corpus of newly-published press releases to recommend to individual readers, where the selected news items should match the reader's reading preference as much as possible. This issue refers to *personalized news recommendation*. Recently, personalized news recommendation has become a promising research direction as the Internet provides fast access to real-time information from multiple sources around the world. Existing personalized news recommendation systems strive to adapt their services to individual users by virtue of both user and news content information. A variety of techniques have been proposed to tackle personalized news recommendation, including content-based, collaborative filtering systems and hybrid versions of these two. In this paper, we provide a comprehensive investigation of existing personalized news recommenders. We discuss several essential issues underlying the problem of personalized news recommendation, and explore possible solutions for performance improvement. Further, we provide an empirical study on a collection of news articles obtained from various news websites, and evaluate the effect of different factors for personalized news recommendation. We hope our discussion and exploration would provide insights for researchers who are interested in personalized news recommendation.

Keywords news recommendation, personalization, scalability, user profiling, modeling, ranking

1 Introduction

Web-based news reading services, such as Google News and Yahoo! News, have become increasingly prevalent as the Internet provides fast access to news articles from various information sources around the world[1]. With the growth of a gigantic number of news articles, a key issue of on-line news services is how to help on-line readers find interesting articles that match the readers' preference as much as possible. This is the problem of *personalized news recommendation*.

In general, given an on-line news reader, the reader's profile information is initially collected by news recommenders to describe his reading preference, and then specific news articles are selected from newly-published press releases to satisfy the reader's reading preference. Traditional approaches for addressing personalized news recommendation involve content-based, collaborative filtering systems, and hybrid versions of these two techniques. The content-based methods enable a recommender system to refine news articles by simply matching the user's reading preference, while the collaborative filtering recommenders aim to analyze reading histories of different users, and then recommend news articles by making use of similar access patterns. Also, hybrid approaches are proposed to alleviate the weakness of individual methods by combining different recommendation techniques[2].

Despite a few recent advances, personalized news recommendation remains challenging for at least three reasons. First, the large number of on-line news articles require the *scalability* of news recommender systems; second, it is not trivial to capture the exact *reading preference* of individual users, since the user's interest evolves over time; third, the *popularity* and *recency* of news articles change dramatically along with the time,

Regular Paper

This work is partially supported by the National Science Foundation of US under Grant Nos. IIS-0546280 and CCF-0830659, and the National Natural Science Foundation of China under Grant No. 61070151.

*Part of the work was done when the author was employed by DailyMe, Inc., Hollywood, FL, 33021, United States.

**Corresponding Author

which differentiates other news items, such as products and movies, rendering traditional recommendation methods ineffective.

In this paper, we provide a comprehensive investigation of the core issues in building an effective personalized news recommender, e.g., how to handle large numbers of news articles, how to effectively utilize different information resources available in news articles to construct users' profiles, and how to select and rank news items. In addition, an empirical study is provided, where several existing news recommendation systems (e.g., content-based methods, collaborative filtering-based methods and hybrid approaches), along with some possible solutions to improving personalized news recommendation, are experimentally investigated, and different factors that might affect the performance of personalized news recommendation are examined. Also, some valuable empirical insights are presented, which can be useful to researchers interested in personalized news recommendation.

The rest of this paper is organized as follows. Section 2 provides a brief review of prior work relevant to personalized news recommendation. To clarify our concerns in this paper, we discuss some core issues of personalized news recommendation in Section 3. To handle the large scale news recommendation problem, several possible solutions are introduced in Section 4. In Section 5 we examine different information resources that can be utilized to build the user's profile. Section 6 presents different strategies of selecting and arranging news items. An empirical study is provided in Section 7 to investigate the effect of different factors in personalized news recommendation. Finally Section 8 concludes the paper.

2 A Review of Personalized News Recommendation

Recently, recommending news articles based on the user's preference has attracted the attention of more and more researchers. Several adaptive news recommending systems, such as Google News and Yahoo! News, provide personalized news recommendation services for a substantial amount of on-line readers. From the methodology perspective, the methods used in the existing news recommendation systems can be broadly categorized into three different groups: content-based methods, collaborative filtering and hybrid approaches.

Content-Based Methods. The content-based approach has been applied to provide personalized selection of news articles in various forms[1,3-7]. In content-based news recommendation systems, news content is being considered when calculating pairwise similarities. Given a set of newly-published news articles and a user with his/her reading history, content-based systems try to sequentially find articles with the content of which matching the user's reading history. Generally speaking, news content is often represented by using vector space model (e.g., Term Frequency-Inverse Document Frequency (TF-IDF)[8]), or topic distributions obtained by language models (e.g., *Probabilistic Latent Semantic Indexing* (PLSI)[9] and *Latent Dirichlet Allocation* (LDA)[10]), and specific similarity measurements are adopted to evaluate the relatedness between news articles. A representative example of such systems involves News Dude[11], a personal news recommending agent that utilizes TF-IDF combined with the K-Nearest Neighbor algorithm to recommend news items to individual users. Another content-based example is YourNews[12], which intends to increase the transparency of adapted news delivery by allowing the given user to adapt his/her profile information. Content-based recommender systems are easy to implement; however, in some scenarios, simply representing the user's profile information by a bag of words is insufficient to capture the exact reading interest of the user.

Collaborative Filtering. Collaborative filtering systems make use of item ratings by users to provide recommendation services, and in general, they are content-free. Particularly for personalized news recommendation, each news article is regarded as an item, and news readers provide item ratings for each article. Here, item ratings are typically binary; a click on a piece of news corresponds to a "1" rating, which indicates that the user is interested in the article, whereas a non-click is represented as a "0" rating[13]. In practice, most collaborative filtering systems are constructed based on users' past rating behaviors, either using a group of users "similar" to the given user to predict news ratings[14-16], or modeling users' behaviors in a probabilistic way[17-19]. Collaborative filtering systems can efficiently capture users' behaviors in case where the overlap in historical consumption across users is relatively high and the content universe is almost static[20]. However, in many web-based scenarios, the content universe undergoes frequent changes, with content popularity changing over time as well[21]. Moreover, many online users are likely to be entirely new with no historical consumption record, which is known as a *cold-start* problem[22]. These issues render collaborative filtering inefficient.

Hybrid Recommenders. As discussed above, content-based and collaborative filtering systems can provide meaningful recommendation but also have some disadvantages. To obtain more reasonable results, many researchers investigate the feasibility of combining these

two types of methods, and propose hybrid solutions to personalized news recommendation. Representative examples include [2, 23], in which the inability of the collaborative filtering to recommend news items is commonly alleviated by using content-based filtering to solve the cold-start problem.

In our previous work[24], we proposed a scalable two-stage personalized news recommendation approach with a two-level representation, which is essentially a hybrid recommender that incorporates content analysis and collaborative filtering, and uses a novel selection strategy to recommend news articles. In this paper, we investigate a couple of key issues to be addressed in personalized news recommendation, and provide a systematic study of how distinct factors can influence the recommendation result. In the following section, we provide an in-depth discussion on the core issues in personalized news recommendation.

3 Issues in Personalized News Recommendation

In general, people prefer instance access to fresh news events. Traditional paper-format releases cannot fulfill such a requirement. An elegant solution is to review news articles via the Internet. When surfing on the Internet, a natural question that a user may face is how to find interesting news events among a myriad amount of news articles. To address such an issue, several popular web-based news reading services, such as Google News and Yahoo! News, have emerged on the Internet to provide news recommendation for many online news readers. Typically, news reading services retrieve news articles relevant to reading preferences of individual users, and adapt their services based on the change of the users' reading interests by employing different recommendation approaches.

However, the exclusive properties of news articles, e.g., unstructured format and short shelf lives, differentiate news recommendation from the ones for other types of Web objects, such as products, movies and people. In the following, we list some unique characteristics of news items.

• *Large Volume.* Different from other types of Web objects, news articles tend to be in flood within a short period of time, requiring much more computation for recommendation.

• *Unstructured Format.* The unstructured format of a news story is more difficult to analyze than other objects with structured properties, e.g., products and friends.

• *Recency.* News items typically have short shelf lives. The majority of stories are consumed within 24 hours of release after which they are often "stale". For instance, few sports fans will care about the boxing scores of two days ago. In contrast, the shelf lives of products and movies extend several months or even years.

• *Entity Preference.* Most news articles describe the occurring of specific events. Online news readers tend to show more interests in the information of what happened, when it happened, where it was and who involved, which are also called named entities.

• *News Selection and Ranking.* The interestedness of news articles with respect to a user is regressive, i.e., after he/she clicks the first piece of news he/she is interested in, the interest value might decrease when he/she clicks the second one or more. Therefore, the ranking of news items recommended to the user deserves careful consideration to maximize the satisfaction of the user.

With these special characteristics of news articles, we summarize the issues of personalized news recommendation as follows.

Scalability. The scalability of news recommendation requires elegant algorithms to efficiently deal with large news corpus. Many strategies can be used to address the scalability issue. For example, *Locality Sensitive Hashing* (LSH)[25] provides an efficient solution to the near neighbor search in high dimensional space by performing probabilistic dimension reduction. *Map-Reduce*[26], a programming model proposed by Google, aims to support distributed computing on large datasets with clusters of computers, and has been widely used in many data mining and machine learning tasks[27-29].

User Profiling. High-quality user profiles can provide thorough representation to describe the user's reading interest, and are very helpful to filter news articles with respect to a given user. Typically, elaborative analysis on unstructured news articles is required for user profiling, which aims to capture the user's reading preference. Particularly, online news readers have special preferences for brief news representations, e.g., what happened, when it happened, where it happened and who involved. User profiling might be facilitated via in-depth analysis of such information.

News Selection and Ranking. Theoretical modeling on news recommendation can provide solutions to the problem of how to select news items that the user might be interested in. Generally speaking, recommendation models are generated based on both news articles and the user's profile, in a way that the personalization is maximized. Further, some other properties, e.g., the *order* of the recommended news list, the *diversity* of the selected news items, are also important indicators for an excellent news recommender.

Results Presentation. Elegant presentation of the

recommended news articles can make online news readers more enjoyable when reading news articles. The recommended news articles are usually presented as a *ranking list*, with extracted snippets to briefly describe news content. One research direction is how to make the selected news items more *diverse*, so that people can acquire more information with distinct topics. In addition, the major idea of a news article may not be precisely expressed by an extracted snippet. An alternative presentation of a single news item involves using *document summarization* techniques[30] to create a concise and informative *summary*. Further, *visualization* techniques can also be employed to make the recommended result more attractive.

In the following, we will delve into the landscape of existing and possible news recommendation systems, to explore how the first three issues are solved in different systems. In addition, we provide valuable insights on different research directions of personalized news recommendation, which can be explored further in future.

4 Large Scale Handling

With the information explosion, it is urgent to deal with large scale news recommendation since fast response of news recommenders is required. To handle this issue, a lot of research efforts have been made in the last decade. In this section, we investigate some existing approaches that can efficiently fulfill the large scale news recommendations; besides, a possible solution based on clustering of news articles is discussed, which might be competent to the large scale task.

4.1 Clustering on Users

A representative system that is capable of dealing with the large scale recommendation is Google News, a web-based news service that generates personalized recommendations to a huge number of online users. In order to handle the large scale problem, a collaborative filtering method[13] is proposed by Google News, in which the underlying relations among different users are explored. Specifically, Minhash clustering technique is first adopted to find users with similar access patterns, and the relationship between users and items is then explored by modeling the joint distribution of users and items as a mixture distribution using PLSI. The interestedness of a news article with respect to a given user is represented as a weighted combination of different scores obtained from Minhash clustering and PLSI. And news candidates are sequentially selected in the order of the interestedness of news articles.

Google News tackles the large scale problem by using probabilistic approximation along with the Map-Reduce framework. One disadvantage of this method is that it mainly depends on the log analysis of the users' access histories. Therefore, it may face the "cold start" problem.

4.2 Clustering on News Articles

An alternative solution to large scale news recommendation involves applying existing techniques to news content, but not to the users. In the domain of news recommendation, news articles contain abundant semantic information of the events they are reporting, i.e., they may belong to different topic categories, which deserves to be utilized for recommendation. In this subsection, we provide a preliminary study of how to employ techniques like LSH to generate an elegant representation of the original news corpus, so that the recommender systems can quickly navigate to the news items that the user might be interested in. This approach can be applied to most content-based news recommender systems.

It is trivial to compute the feature spaces and compare all pairs of articles in the set of hundreds of news articles; however, such pairwise comparisons become computationally expensive for a larger news corpus. For instance, a news corpus with 100 000 news items requires ten billion pairwise comparisons. In reality, news articles describing widely different stories do not have enough distinguishable words in common, which renders the feature space more sparse. Therefore, the traditional pairwise comparison-based methods are inefficient for large news datasets.

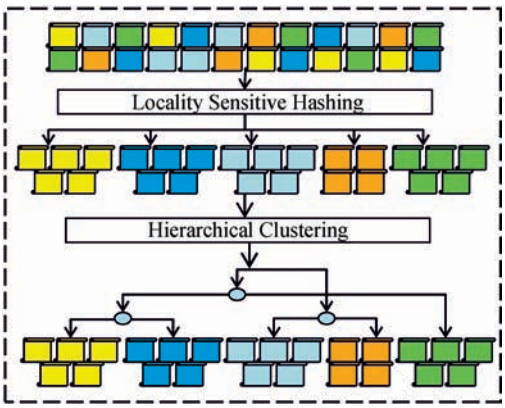


Fig.1. Clustering on news articles.

To tackle this problem, LSH can be utilized to eliminate unnecessary similarity computations between unrelated articles, and obtain a rough separation on the original news corpus. After separating news corpus into small groups, we can employ hierarchical clustering to arrange these groups into intermediate clusters. The original news corpus is then represented as a two-level

news hierarchy, where the first level contains several intermediate clusters, and the second level lists specific news articles within each cluster. With such representation, one can quickly find specific news articles by accessing the generated news hierarchy from top to bottom. Fig.1 shows a brief framework of clustering on news articles.

4.3 Discussion on Scalability Issue

Recall that the ultimate goal of news recommendation is to provide news articles to individual users. Users may be concerned about how interesting the recommended news articles are. By following the above framework, we can choose different news articles that reside in distinct intermediate news clusters by navigating through the two-level news hierarchy. The recommended news items will carry diverse information to broaden the user's reading scope. From this perspective, clustering on news articles is more reasonable than clustering on users in terms of digesting news content.

Clustering on news articles aims to separate the original news set into multiple clusters. "LSH + hierarchical" clustering can be used to fulfill this purpose: LSH on news articles and hierarchical clustering on news groups. Alternative ways of grouping news articles involve the standard K-means and hierarchical clustering directly on the original news corpus. However, these two clustering techniques are not efficient when dealing with large text corpus. Also, when recommending news items, we cannot quickly find a small news group that the user might be interested in by using these two clustering techniques.

Further, we choose hierarchical clustering method to separate news groups instead of using other clustering approaches, like K-means. The intuitive reason of using hierarchical clustering is that hierarchical clustering provides us an elegant global representation of the latent structure of news corpus, which is helpful to the subsequent procedures, e.g., selecting news articles to recommend to individual users.

5 User Profiling

Personalized news recommendation starts with constructing a user's profile by analyzing the reading history of the user. In general, there are several types of resources available in the user's reading history. Traditional explorations of user profiling ranged from analyzing news content to making use of similar access patterns of different users. A comprehensive survey of various user profile construction techniques is provided in [31]. In the following, we discuss some general approaches to build users' profiles, along with preliminary attempts on the usage of different resources available in users' reading histories.

5.1 Content-Based Profiling

In general, content-based profiling aims to build users' profiles by virtue of news content. News content can be represented by using either term weighting methods or concept weighting approaches.

5.1.1 Term Weighting

A well-known term weighting method is TF-IDF (term frequency-inverse document frequency)[8]. As discussed in [12], before calculating the TF-IDF values, a sequence of preprocessing steps are executed, including removing stop words, tokenizing, stemming and so on. Then news articles that the user has read in the past will be quantified as a term vector stored in the user's profile, where each entry is the TF-IDF value of the corresponding term. For recommendation, a natural way is to compute the pairwise similarities between the user's profile and the newly-published news items, by employing appropriate similarity measurements, e.g., the cosine similarity[32]. Then news articles are sequentially selected in the order of similarity ranking. Term-weighting-based profiling is easy to understand and implement; however, the high dimensionality of the generated vector space renders personalized news recommendation computationally intractable for large scale news corpus.

5.1.2 Concept Weighting

In traditional forms of news content comparison, all words contained in news articles are considered. In addition to this, there are no explicit relation between different terms[33]. For instance, it is difficult to determine the relation between Google and Yahoo! by using term weighting. However, an online news reader who is interested in news regarding his stocks in Google, might also prefer to read news about Yahoo!, since it is a competitor of Google. To handle the insufficiency of term weighting, several researchers propose to adopt the "ontology" to discover the implicit semantic relations between different terms, and then quantify each article as a weighted concept vector, where the concepts are derived from semantically related terms. For example, [32] presents a semantic-based recommendation system, in which the user's profile is represented as weighted concept vectors, and several similarity measurements are adopted to calculate the pairwise similarities between articles. Obviously, the resulted concept vector is with much smaller dimensionality than the term weighting vector, and therefore the computation of pairwise article similarities becomes less expensive.

5.2 Collaborative-Based Profiling

In many real-world scenarios, online news readers may exhibit similar reading preference. For users within a social group, their reading behaviors can be easily influenced by their friends in an explicit way; comparatively, users who read similar news articles may compose an implicit reading group, even if they do not know each other, which means it is transparent to users. A user's profile information can be enriched by analyzing other users' reading preference similar to that of the given user, which is essentially collaborative filtering (CF).

When analyzing accessed news data, there are a huge number of users that have frequent online reading records. Some users may share similar access patterns with other users. To discover such information, a natural way is to employ clustering techniques on users. Clustering users could reduce the data dimension and may also benefit collaborative filtering in future. Since we have the user profile, a user-category matrix could be constructed. For example, each row of the matrix is a user, and each column is a category, and the value is the percentage of the information in the news articles, that the user has read, belonging to the category. Once we have the matrix, clustering algorithms such as K-means can be applied to obtain the user groups. Another more comprehensive analysis on the user clustering would be transferring information of news articles (e.g., the news category information) to help the user grouping. If we have the user clusters, the nearest neighbors of a user can be easily detected within the cluster he/she belongs to, and therefore the user's profile can be represented as a group of users that have similar access patterns with respect to the given user.

5.3 Entity-Based Profiling

Typically, in news articles, named entities include information describing what happened, when the event happened, where it happened, who were involved, and so on. In general, named entities contain much more semantic information and relations than simple terms. News readers might have special preference for some particular named entities contained in news articles. For example, when an online news reader is reading an article related to "*Haiti Earthquake*", what the user cares about may include when the earthquake happened, how many people were affected, but not detailed descriptive snippets. Therefore, *named entities* are important when building users' profiles to capture the exact reading preference. However, little research efforts have been made to explore the effect of using named entities for personalized news recommendation.

Similar to the content-based profiling, news articles can be represented as an entity vector that describes what entities are involved in the articles. Such a vector can also be treated as the user's profile. To extract named entities, some open source natural language processing tools are available on the Internet. For example, GATE[34], a software package for text processing, is capable of automatically identifying named entities in texts by predefining a couple of entity rules. By default, we can use the rules oriented from GATE. After entity detection, each news article is associated with a list of named entities along with their corresponding entity types, and the user's profile can then be represented as an integrated entity vector by combining named entities of accessed news articles.

5.4 Hybrid Profiling

An intuitive observation of user profiling is that if the user's profile contains more information, then it is much closer to the user's exact reading preference. In other words, the weakness of individual recommendation methods can be alleviated by combining them together. To encapsule more information into users' profiles, many researchers propose to combine different kinds of resources that we introduced above, to generate hybrid solutions to user profiling. For example, Burke[2] provides a comprehensive study that examines a large subset of the hybrid recommendation design space. However, there is little effort that integrates the entity-based method into the hybrid solutions. In the experiment, we provide an empirical comparison between entity-involved methods and nonentity-involved methods.

5.5 Users' Interest Evolution

Another remarkable point in personalized news recommendation is the interest evolution of news readers. Typically, the user's reading preference changes over time. By capturing the evolutionary interest trend of users, we can easily reveal what is the concern of most users, and how users are linked by real-time press releases, which deserves to be explored in future research.

6 News Selection and Ranking

After obtaining the user's profile, the subsequent procedure is to select and rank news articles that match the user's reading preference as much as possible. In regular, the user's reading interest is covered in the user's profile, e.g., what topics that the user might be interested in, who have the similar reading interests with the user, and what kinds of entities the user might prefer. Given such information, traditional

recommendation algorithms employ greedy algorithms based on the pairwise similarity between the user's profile and the article being selected. However, due to the special characteristics of news articles (e.g., the popularity and recency), it is not trivial to come up with a reasonable and preferable recommendation list.

6.1 News Selection

In most recommender systems, greedy algorithms are adopted to recommend news articles to individual users, assuming that news articles are independent each other. However, in most scenarios, the intra relations among different news articles render them dependent in terms of news recommendation. For example, given two news articles that talk about the actor cast of a popular movie "*Inception*", online news readers might be likely to click only one of them, and discard another. In addition, when the users are selecting news articles to read, the interestedness of news articles with respect to a user is regressive, i.e., users tend to read the most interesting one first, and subsequently select the less attractive ones. Therefore, a deep exploration of the intra relations among different news articles might be helpful to capture the user's preference, thus provide high-quality news recommendation results.

6.1.1 Bandit Modeling

Besides the general greedy algorithm, there are many other ways to select news items, depending on how to model the selection procedure. One possible solution to this issue is provided in [21], in which personalized recommendation of news articles is modeled as a contextual bandit problem, where a learning algorithm sequentially selects articles to serve users based on contextual information about users and articles, while simultaneously adapting its article-selection strategy based on user-click feedback to maximize total user clicks. Essentially, their method tries to address the issue of optimally balancing the two competing goals: maximizing user satisfaction in the long run, and gathering information about goodness of match between user interests and content.

6.1.2 Submodularity Modeling

Another possible solution involves utilizing the submodularity[35] hidden in various news articles to model personalized news recommendation problem. In reality, some news articles concentrate on similar or even the same topic, with minor difference on major aspects of the corresponding topic. For example, given a news group talking about the movie "*Inception*", one piece of news may focus on the actor cast of this movie, while another may describe the high-end techniques used in this movie. Typically, a news reader is interested in some specific aspects of the given topic, but not all of them. In addition, the interestedness of news articles with respect to a specific user is regressive, i.e., the user always picks the most interesting article to read. With such diminishing property, personalized news recommendation can be modeled as a budgeted maximum coverage problem[36], and be solved by employing greedy algorithms[24].

Furthermore, submodularity-based modeling can be extended to different directions, but not restricted in the previous discussion. In essence, submodularity aims at modeling users' reading behaviors in a premise that a specific factor decreases when the user reads more news articles. For example, such a factor can be the reading time: 1) time spent on each successive article in the same category/cluster is progressively lower, and 2) the same article is read for longer time if it appears earlier in the reading path for a user. Therefore, submodularity-based modeling is more flexible and effective when modeling user's reading behaviors, and consequently can provide a more reasonable reading list of news articles for individual users.

6.1.3 Semi-Supervised Modeling

Most news recommender systems focus on analyzing the user's reading history to get a general understanding of the user's preference. In real-world applications, recommendation is usually fulfilled by unsupervised methods, involving clustering, similarity matching and so on. However, the natural property of news articles that the user has read in the past provides valuable categorical information (read or not read by users), which can be regarded as "labels" of news articles. A click on a piece of news corresponds to a "read" label, whereas a non-click is represented as a "noread" label. Based on this observation, one can easily model personalized news recommendation as a semi-supervised learning problem, i.e., to predict the labels of newly-published news articles with respect to a given user, and then choose the articles with the "read" label as the recommendation result. This semi-supervised modeling is also interesting in terms of news recommendation.

6.2 News Ranking

Recommendation modeling helps to select news articles, but not ranking them. In fact, the ranking of the recommended news items can be influenced by several factors, such as the exclusive properties of news articles, the reading preference of the given user, and even the preference of other users who have similar access patterns with the given user. These implicit factors should

be considered when ranking news articles.

6.2.1 Property-Based Ranking

As discussed in Section 3, news stories are different from other web objects, such as products and movies, in terms of the popularity and the recency. In general, these two characteristics are temporally in effect with respect to news articles. When a news article is published online, there might be a lot of news readers to click it, rendering this article very popular; however, as the time evolves, the interestedness in the news article might decrease, since most online users have their reading preference on newly-published press releases. Therefore, when recommending news articles to individual news, these two factors should be taken into consideration, so that the resulting news list will capture more instant news events as well as satisfy the user's reading preference.

6.2.2 Preference-Based Ranking

A user's reading preference is an important factor that describes what kind of news articles this user might be interested in. With the outstanding progress of information diversity, most news readers are likely to read press releases within different topic categories, such as sports, movies, and so on. An elegant way to capture the user's preference for different topic categories is to represent the user's reading interest as a weighted distribution of different topics. Then based on the topic distribution, we can rank the selected news articles by following the order of the topics scores.

To get the topic distribution of the user's reading history, one can employ probabilistic language models, such as PLSI[9] and LDA[10]. The PLSI model and the LDA model are similar in terms of probabilistic language models, except that in LDA the topic distribution is assumed to have a *Dirichlet prior*. Note that the PLSI model is equivalent to the LDA model under a uniform *Dirichlet prior* information, whereas the LDA model is essentially the *Bayesian* version of the PLSI model[37]. *Bayesian* formulation tends to perform better in small datasets because *Bayesian* methods can avoid overfitting. In a very large dataset, the results are probably the same. One difference is that PLSI uses a variable d to represent a document in the training set. As discussed in [10], when a model representing a document has not been seen before, PLSI fixes the probability of words under topics to be that learned from the training set and uses the same Expectation-Maximization (EM) algorithm to infer the topic distribution under a document. Blei argues that this step is cheating because the model is essentially refitted to the new data[10].

6.2.3 Group-Based Ranking

In reality, many online news readers may exhibit similar reading preference. Recently, the tremendous growth of online social network services greatly facilitates the collaboration, sharing, and other interactions among on-line users. For users within a social group, their reading behaviors can be easily influenced by their friends in an explicit way; comparatively, users who read similar news articles may compose an implicit reading group, even if they do not know each other or they do not realize this happens. Here, if a user reads an articles, then we can regard it as a vote on this article. When recommending news items to a given user, the ranking of the selected news pieces can be determined by the number of votes on these articles cast by other users who have similar access patterns with the given user.

7 An Experimental Study

In this section, we provide an empirical exploration of the issues discussed above. We start with an introduction to a real-world news collection obtained from multiple news service websites. Then we design a set of experiments to empirically evaluate how different factors affect the performance of personalized news recommenders, and shed light on potential research directions for researchers interested in personalized news recommendation.

Table 1. Statistical Description of News Dataset

Category	No. Articles
Economy	18 204
Entertainment	20 460
Environment	4 147
Health	5 542
Law	11 873
Politics	21 240
Science	8 815
Sports	9 562
World	12 537

7.1 Real World Dataset

For experiments, we used news articles along with users' access history from two popular news websites (see details in [24]). Both websites contain multiple news topic categories, such as sports, movies and politics. We used the news data for 9 categories on purpose, where the data collection ranged from Aug. 15th, 2010 to Nov. 16th, 2010. For each news article, we queried the online readers who had read it, and then recorded readers' information in an anonymous way. After obtaining the whole dataset, we preprocessed the data by

removing news articles that were rarely accessed (i.e., the accessed frequency is less than 10 times per day) and by storing users with frequent online reading behaviors (i.e., users who read news articles every day and read more than 10 pieces of news each day). After preprocessing, a total of 112 380 news items were stored, each day in average 1 221 news articles, along with 4 630 users. A brief statistical description of the news collection is presented in Table 1.

7.2 Clustering on News Articles

To validate the feasibility of LSH-based clustering, we selected several time ranges with different intervals and ran the clustering procedure on news articles from these time ranges. For the purpose of comparison, we also implemented the standard K-means① and hierarchical clustering with average-link, to compare the F-measure and the time cost. All the clustering algorithms were implemented using Java and tested under the same experimental environment. For K-means and hierarchical clustering, we represented each news article using vector space model, each entry of which was denoted by TF-IDF value of the corresponding term. In each time range, we executed different algorithms 10 times respectively, except hierarchical clustering, and averaged the F-scores[32] as the final results. Note that the number of clusters is set to be the number of categories we introduced in Subsection 7.1.

We used the standard $F1$ measure[32] to evaluate the accuracy of clustering results. To evaluate the average performance across multiple clusters with different cardinality, the micro-averaging $F1$ and macro-averaging $F1$[32] were adopted. The former gives equal weight to every instance, and it tends to be dominated by large size clusters; the latter assigns equal weight to each cluster and it is influenced by clusters with a relatively small number of instances.

Table 2 lists the clustering evaluation results. Based on the comparison, we observed that:

1) The "LSH + Hierarchical" clustering on news articles significantly outperforms the general K-means and hierarchical clustering techniques in terms of accuracy and efficiency. A straightforward explanation of the accuracy increase is that we used *shingles* as the representation of news articles. Shingling aims to separate articles into shingles where the probability of any given shingle appearing in any article is low. In this way, similar articles will have more shingles in common, whereas dissimilar articles share few shingles.

2) The accuracy of K-means and hierarchical clustering on news corpus decreases when the scalability of the dataset enlarges.

Therefore, "LSH + Hierarchical" clustering can be a valid solution to addressing personalized recommendation of large volume of news articles.

7.3 User Profiling

For user profiling, we tested the effect of each possible combination of the three factors (i.e., news content, preferred entities and collaborative information) to the recommendation results. Here we used a greedy algorithm to select news articles similar to the user's profile. Specifically, we selected 100 users from the users' pool, randomly picked up 10 time ranges with 3-day interval from our news dataset, and recommended news items (top @10, @20 and @30) to these users based on different aspect combinations. For comparison, we computed the averaged F-score of recommendation results for the 100 selected users over 10 time ranges. Fig.2 depicts the result.

From the comparison, we observed that:

1) The hybrid approach that *considers all the three aspects* always performs the best.

Table 2. "LSH + Hierarchical" Clustering vs Direct Clustering

Time Range	No. Articles	K-Means			Hierarchical (average-link)			"LSH + Hierarchical"		
		Micro-F1	Macro-F1	Time Cost (min)	Micro-F1	Macro-F1	Time Cost (min)	Micro-F1	Macro-F1	Time Cost (min)
08/15∼08/16	2 340	0.453 0	0.374 4	3	0.439 8	0.355 1	5	0.501 6	0.462 2	2
08/19∼08/22	5 572	0.436 5	0.365 3	5	0.420 7	0.350 6	9	0.489 7	0.440 6	4
08/23∼08/31	10 985	0.422 7	0.342 1	10	0.415 6	0.341 2	15	0.490 2	0.453 5	7
09/01∼09/15	18 841	0.396 2	0.330 2	21	0.382 0	0.340 9	25	0.483 0	0.448 0	10
09/01∼09/30	35 920	0.378 3	0.315 3	38	0.377 7	0.334 4	50	0.512 8	0.459 8	15
10/01∼11/16	63 659	0.352 9	0.297 6	59	0.375 5	0.322 0	78	0.523 9	0.467 9	25
08/15∼11/16	112 380	0.321 7	0.271 1	121	0.362 2	0.311 8	152	0.509 3	0.450 6	38
Average	–	0.394 5	0.328 0	–	0.396 2	0.336 7	–	**0.501 5**	**0.454 7**	–

①In the experiment, the maximum number of iterations for K-means algorithm is set to be 100.

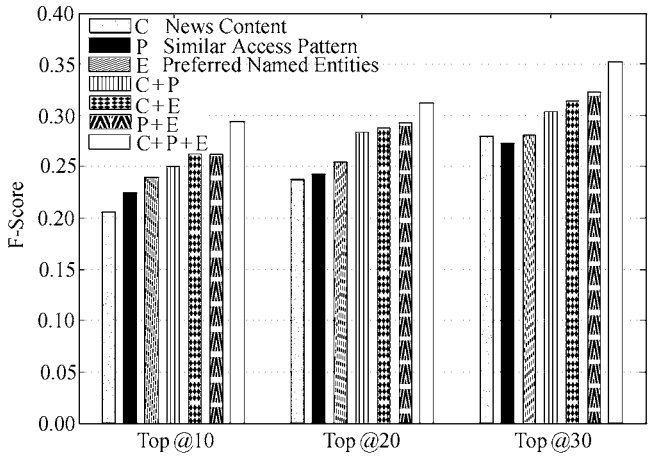


Fig.2. Recommendation F-score of different combinations of profile factors.

2) Recommendation purely *based on one single aspect* cannot achieve as good performance as the system based on the combinations. The reason is straightforward: more intrinsic properties of news articles and users' profiles can be revealed as we take more aspects into account.

3) Recommender systems with *preferred named entities* involved perform better than the ones without considering named entities. This observation verifies our claim that news readers tend to show more interest in simple but representative named entities contained in news articles.

7.4 Feature Representation

In this subsection, we try to explore the feasibility of using semi-supervised learning to recommend news articles to individual users. Specifically, we generate a sample dataset from the news collection by selecting users who read 40∼60 news articles in the past. Thus, in this scenario, we do not consider users reading too few and too many articles. The dataset contains 1 589 users and 5 953 articles with titles and story bodies. Fig.3 shows the news reading account in this sample dataset.

For each user we use the libSVM classifier[38], a software package that implements support vector machines for classification problems, to predict the labels for the articles. We use 70% of the article data for training and 30% of them for testing. In practice, we have the articles from two classes {read(1), not read(0)}, and we randomly select 70% of the articles in each class and combine them to obtain the training dataset. Similarly we get the testing data. In this case, we avoid the situation of all the training (or testing) data belonging to only one class. Once we obtain the classification results, the articles with label "1" will be recommended to the user.

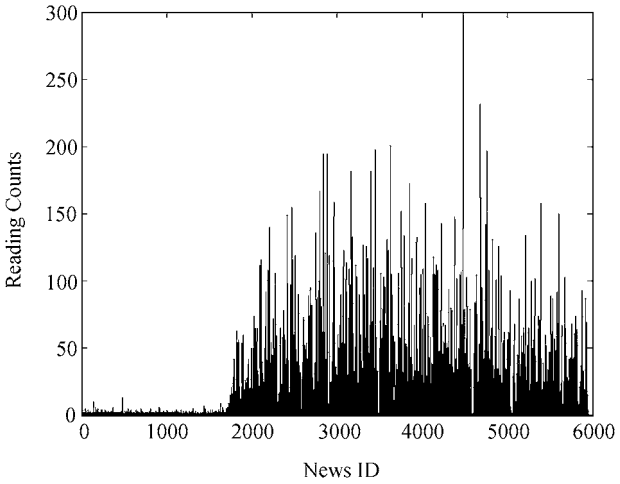


Fig.3. Statistics of the number of times that news articles have been read.

In the experiment, we compared different feature generation methods as follows.

• All: uses all the terms as the features;

• V1000: uses the top 1 000 frequent terms as the features;

• FS100: uses a feature selection algorithm (maximal relevance and minimal redundancy)[32] to select the top 100 features from the V1000;

• FS50: similar to FS100, uses top 50 selected features;

• FS20: uses top 20 selected features.

Based on the characteristics of the data, two classes are extremely unbalanced, e.g., only a few articles that we need to recommend to users, and a large volume of them are not in the recommendation list. Therefore, if we evaluate the results using accuracy as in traditional classification evaluation, in the extreme case, even if all the news are labeled with "0", the accuracy is still very high. However, we care more about the articles with the label "1". To handle this issue, precision and recall for the class "1" are used as the evaluation metrics in the experiments. The definitions of precision and recall are as follows:

$$precision = \frac{|\{history\} \cap \{recom\}|}{|\{recom\}|}, \tag{1}$$

$$recall = \frac{|\{history\} \cap \{recom\}|}{|\{history\}|}. \tag{2}$$

Table 3 shows the average results for all the users

using SVM based on different feature sets. From the results, we have the following observations.

Table 3. Results by Using Different Feature Sets

	All	V1000	FS100	FS50	FS20
Precision	61%	78%	79%	76%	83%
Recall	9%	21%	29%	29%	56%

• Precision scores are higher than recall scores because in many cases only a few articles are assigned to the label "1".

• If we use all the terms in the articles as the features, the results are very poor due to the *sparseness* and *noise* in the data.

• Using *frequent terms* to build a vocabulary can improve the performance of the classification.

• *Feature selection* methods do help the classification because a more robust learning model can be built based on *relevant features*.

• The number of selected features needs to be decided, and in this case when we choose the top 20 features, the result outperforms all the others.

Beyond this, we also compare the effect of using term features and entity features. Fig.4 shows the recall scores of four implemented methods: 1) classification using top 500 frequent terms as features; 2) classification for one cluster of users (clustering users into 10 groups using K-means on the user-articles matrix) using top 500 term features; 3) classification using top 500 frequent entities as features; 4) classification for the same cluster of users (using the same clustering method) using top 500 entity features.

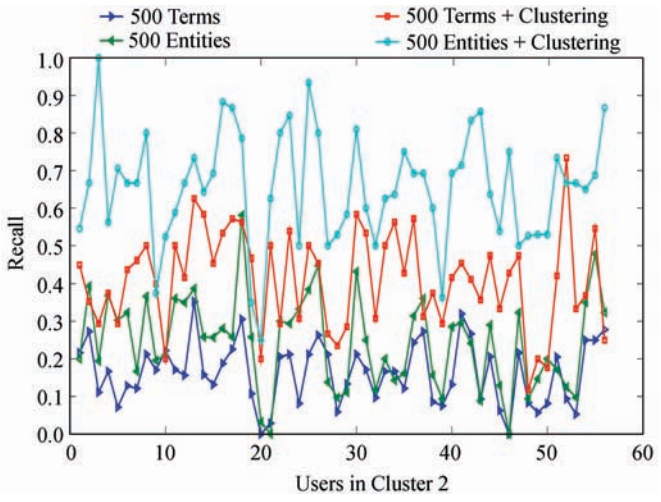


Fig.4. Recall scores for each user in one example cluster using the four methods.

From the results, we have the following observations.

1) *Entity features outperform term features.* This is not surprising because entities extracted from texts contain more semantic information and relations than terms.

2) It is not fair to conclude that user clustering will benefit recommendation using the recall scores for individual users, because in the clustering scenario, we obtain more articles with the label "1" than in each individual classification model for each user. However, the recall scores for the whole group of users are 34% and 45% using terms and entities, respectively. The average recall scores returned from the individual models are 16% and 24% using terms and entities, respectively. This means *the model for the whole cluster of users is better than the average models for the individual users in the cluster*, and the reasons may lay on the facts that the data are more balanced and the group of users share some common interest so that the reading patterns are easier to find.

8 Conclusion

In this paper, we provide a comprehensive investigation of existing and possible news recommenders, and discuss some core issues in news recommendation. Personalized news recommendation of large volume of news articles needs to be addressed by employing efficient processing techniques. Also, to better capture online news readers' exact reading preference, we explore different user profiling approaches, and provide comprehensive comparisons from both the theoretical and practical perspectives. Furthermore, several insightful recommendation models, along with important ranking factors, are discussed in this paper. We also provide an empirical studies on a collection of news articles obtained from various news websites. We hope our discussion would be helpful to researchers interested in personalized news recommendation.

Acknowledgement We would like to acknowledge DailyMe, Inc. for sharing data with us and providing useful comments and suggestions, especially the following individuals in the DailyMe team: Dr. Balaji Padmanabhan, Daniel Knox, Jose Zozaya, Eduardo Hauser, and Neil Budde.

References

[1] Liu J, Dolan P, Pedersen E R. Personalized news recommendation based on click behavior. In *Proc. the 14th International Conference on Intelligent User Interfaces*, Hong Kong, China, Feb. 7-10, 2010, pp.31-40.

[2] Burke R. Hybrid systems for personalized recommendations. In *Proc. Workshop on Intelligent Techniques for Web Personalization*, Acapulco, Mexico, Aug. 11, 2005, pp.133-152.

[3] Billsus D, Pazzani M J. User modeling for adaptive news access. *User Modeling and User-Adapted Interaction*, 2000,

10(2): 147-180.

[4] Carreira R, Crato J M, Gonçalves D, Jorge J A. Evaluating adaptive user profiles for news classification. In *Proc. the 9th International Conference on Intelligent User Interfaces*, Funchal, Brtngal, Jan. 13-16, 2004, pp.206-212.

[5] Kim H R, Chan P K. Learning implicit user interest hierarchy for context in personalization. *Applied Intelligence*, 2008, 28(2): 153-166.

[6] Liang T P, Lai H J. Discovering user interests from web browsing behavior: An application to internet news services. In *Proc. HICSS*, Hawaii, USA, Jan. 7-10, 2002, pp.2718-2727.

[7] Tan A H, Teo C. Learning user profiles for personalized information dissemination. In *Proc. IEEE International Joint Conference on Computational Intelligence*, Horolulu, USA, May 12-17, 2002, pp.183-188.

[8] Jurafsky D, Martin J H, Kehler A, Vander Linden K, Ward N. Speech and Language Processing. Prentice Hall, 2000.

[9] Hofmann T. Probabilistic latent semantic indexing. In *Proc. the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, USA, Aug. 15-19, 1999, pp.50-57.

[10] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 2003, 3: 993-1022.

[11] Billsus D, Pazzani M J. A personal news agent that talks, learns and explains. In *Proc. the 3rd Annual Conference on Autonomous Agents*, Seattle, USA, May 1-5, 1999, pp.268-275.

[12] Ahn J, Brusilovsky P, Grady J, He D, Syn S Y. Open user profiles for adaptive news systems: Help or harm? In *Proc. the 16th International Conference on World Wide Web*, Banff, Canada, May 8-12, 2007, pp.11-20.

[13] Das A S, Datar M, Garg A, Rajaram S. Google news personalization: Scalable online collaborative filtering. In *Proc. the 16th International Conference on World Wide Web*, Banff, Canada, May 8-12, 2007, pp.271-280.

[14] Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J. GroupLens: An open architecture for collaborative filtering of netnews. In *Proc. the 1994 ACM Conference on Computer Supported Cooperative Work*, Chapel Hill, USA, Oct. 22-26, 1994 pp.175-186.

[15] Sarwar B, Karypis G, Konstan J, Reidl J. Item-based collaborative filtering recommendation algorithms. In *Proc. the 10th International Conference on World Wide Web*, Hong Kong, China, May 1-5, 2001, pp.285-295.

[16] Yu K, Xu X, Tao J, Ester M, Kriegel H P. Instance selection techniques for memory-based collaborative filtering. In *Proc. the 2nd SIAM International Conference on Data Mining*, Arlington, USA, Apr. 11-13, 2002, pp.59-74.

[17] Breese J S, Heckerman D, Kadie C *et al*. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. the 14th Conference on Uncertainty in Artificial Intelligence*, Madison, USA, Jul. 24-26, 1998, pp.43-52.

[18] Hofmann T. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 2004, 22(1): 89-115.

[19] Shani G, Heckerman D, Brafman R I. An MDP-based recommender system. *Journal of Machine Learning Research*, 2006, 6(2): 1265.

[20] Schafer J B, Konstan J, Riedi J. Recommender systems in e-commerce. In *Proc. the 1st ACM Conference on Electronic Commerce*, Denver, USA, Nov. 3-5, 1999, pp.158-166.

[21] Li L, Chu W, Langford J, Schapire R E. A contextual-bandit approach to personalized news article recommendation. In *Proc. the 19th International Conference on World Wide Web*, Raleigh, USA, Apr. 26-30, 2010, pp.661-670.

[22] Schein A I, Popescul A, Ungar L H, Pennock D M. Methods and metrics for cold-start recommendations. In *Proc. the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tenpere, Finland, Aug. 11-15, 2002, pp.253-260.

[23] Chu W, Park S T. Personalized recommendation on dynamic content using predictive bilinear models. In *Proc. the 18th International Conference on World Wide Web*, Madrid, Spain, Apr. 20-24, 2009, pp.691-700.

[24] Li L, Wang D, Li T, Knox D, Padmanabhan B. SCENE: A scalable two-stage personalized news recommendation system. In *Proc. the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Beijing, China, Jul. 25-29, 2011, pp.124-134.

[25] Gionis A, Indyk P, Motwani R. Similarity search in high dimensions via hashing. In *Proc. the 25th International Conference on Very Large Data Bases*, Edinberg, UK, Sept. 7-10, 1999, pp.518-529.

[26] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 2008, 51(1): 107-113.

[27] Chu C T, Kim S K, Lin Y A, Yu Y Y, Bradski G, Ng A Y, Olukotun K. Map-reduce for machine learning on multicore. In *Proc. the 2006 Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 4-7, 2006, pp.281-288.

[28] Kang U, Tsourakakis C E, Faloutsos C. PEGASUS: A petascale graph mining system implementation and observations. In *Proc. the 9th IEEE International Conference on Data Mining*, Miami, USA, Dec. 6-9, 2009, pp.229-238.

[29] Papadimitriou S, Sun J. Disco: Distributed co-clustering with map-reduce: A case study towards petabyte-scale end-to-end mining. In *Proc. the 8th IEEE International Conference on Data Mining*, Pisa, Italy, Dec. 15-19, 2008, pp.512-521.

[30] Wang D, Zhu S, Li T, Gong Y. Comparative document summarization via discriminative sentence selection. In *Proc. the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, China, Nov. 2-6, 2009, pp.1963-1966.

[31] Gauch S, Speretta M, Chandramouli A, Micarelli A. User Profiles for Personalized Information Access. The Adaptive Web, 2007, pp.54-89.

[32] Tan P N, Steinbach M, Kumar V *et al*. Introduction to Data Mining. Boston: Pearson Addison Wesley, 2006.

[33] IJntema W, Goossen F, Frasincar F, Hogenboom F. Ontology-based news recommendation. In *Proc. the 2010 EDBT Workshops*, Laussane, Switzerland, Mar. 22-26, 2010, pp.1-6.

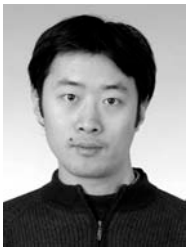
[34] Cunningham D H, Maynard D D, Bontcheva D K, Tablan M V. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. the 40th Anniversary Meeting of the Association for Computational Linguistics*, Philadelphia, USA, Jul. 6-12, 2002, pp.168-175.

[35] Nemhauser G L, Wolsey L A, Fisher M L. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 1978, 14(1): 265-294.

[36] Khuller S, Moss A, Naor J S. The budgeted maximum coverage problem. *Information Processing Letters*, 1999, 70(1): 39-45.

[37] Girolami M, Kabán A. On an equivalence between PLSI and LDA. In *Proc. the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, Jul. 28-Aug. 1, 2003, pp.433-434.

[38] Chang C C, Lin C J. LIBSVM: A library for support vector machines. *ACM Trans. Intelligent Systems and Technology*, 2001, 2(3): Article No.27.



Lei Li received his M.S. degree in software engineering from Beihang University in 2008. He is currently a Ph.D. candidate in School of Computing and Information Sciences at Florida International University. His research interests include data mining and machine learning.

Ding-Ding Wang received her Bachelor's degree from the Department of Computer Science, University of Science and Technology of China in 2003, and her Ph.D. degree in computer science in 2009 from Florida International University. She is currently a postdoctoral researcher in the Center for Computational Science at University of Miami. Her research interests are data mining and information retrieval.

Shun-Zhi Zhu received his Ph.D. degree in control theory and engineering in 2007 from Xiamen University. He is currently an associate professor and vice chair of the Department of Computer Science & Technology at Xiamen University of Technology. His research interests are information systems, GIS, and data mining.

Tao Li received his Ph.D. degree in computer science in 2004 from the University of Rochester. He is currently an associate professor in the School of Computer Science at Florida International University. His research interests are in data mining, machine learning and information retrieval. He is a recipient of USA NSF CAREER Award and multiple IBM Faculty Research Awards.