Incorporating User Behaviors in New Word Detection

Yabin Zheng[1], Zhiyuan Liu[1], Maosong Sun[1], Liyun Ru[1,2], Yang Zhang[2]
[1]State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Sohu Inc. R&D center, Beijing 100084, China
{yabin.zheng, lzy.thu}@gmail.com sms@mail.tsinghua.edu.cn {ruliyun, zhangyang}@sohu-rd.com

Abstract

In this paper, we proposed a novel method to detect new words in domain-specific fields based on user behaviors. First, we select the most representative words from domain-specific lexicon. Then combining with user behaviors, we try to discover the potential experts in this field who use those terminologies frequently. Finally, we make further efforts to identify new words from behaviors of those experts. Words used much more frequently in this community than others are most probably new words. In brief, our method follows a collaborative filtering way: first from words to find professional experts, then from experts to discover new words, which is different from the traditional new word detection methods. Our method achieves up to 0.86 in accuracy on a computer science related data set. Moreover, the proposed method can be easily extended to related words retrieval task. We compare our method with Google Sets and Bayesian Sets. Experiments show that our method and Bayesian Sets gives better results than Google Sets.

1 Introduction

New word detection [Chen and Bai, 1998; Li, *et al.*, 2004; Peng, *et al.*, 2004; Wang *et al.*, 1995; Zhang, *et al.*, 2002] is one of the most important and crucial issues in Chinese natural language processing. In fact, it is quite related to Chinese word segmentation problem. More specifically, [Sproat and Emerson, 2003] shows that out-of-vocabulary words have a crucial impact on Chinese word segmentation task. Meanwhile, with the rapid development of internet technology, new words emerge very frequently. Chinese new word detection is even more challenging because there are no natural word boundaries in Chinese. Previous works [Peng, *et al.*, 2004; Zhang, *et al.*, 2002] indicate that new word detection is better performed concurrently with Chinese word segmentation (CWS) in most cases. Unfortunately, CWS itself is a more challenging task. In this study, we leave CWS out of consideration. By incorporating user behaviors, we perform new word detection task in a different and novel manner. We will present detailed explanations in the following sections.

Recently, a number of methods have been applied to collaborative filtering [Adomavicius and Tuzhilin, 2005; Breese *et al.*, 1998; Nakamura and Abe, 1998; Hofmann, 2003]. Usually, the collaborative filtering systems produce personal recommendations based on the similarities between the preference of the active user and others. Assume that if you need to choose among various items which you have no experience with before. The direct solution is to ask opinions of others who have the similar taste. If an item is liked by most of those users, maybe you are tending to choose it too. Examples of such collaborative filtering applications include recommending movies by MovieLens and Netflix [Miller *et al.*, 2003], books at Amazon [Linden *et al.*, 2003], and so on. In this paper, we borrow the idea of collaborative filtering. The underlying idea of our method is that users on the same professional field tend to have the same taste on certain special words. In other words, they will use those words more frequently than other users. We regard those special words as the candidates of new words. Finally, since there are different definitions of new words under different situations, we specify a clear and concise definition here. New words in this paper refer to the words that should be included in certain lexicon but are not included, which is analogous to the definition of out-of-vocabulary words.

The contributions of this paper are twofold. First, we propose a framework that integrates user behaviors in new word detection task, which is quite different from other methods. To the best of our knowledge, this paper is the first one which introduces user behaviors in new word detection task. Second, we extend our proposed method to related words retrieval task. Given a query consisting of a discriminative word, we are able to return several similar or related words. As shown in the experiment section, our proposed method gains reasonable performance in both tasks.

The rest of the paper is organized as follows. Some related works are discussed in section 2. We mainly focus on works about new word detection, collaborative filtering and Google Sets. Then we will introduce some background, including the dataset we used and some basic concepts like user dictionary and cell dictionary in section 3. After that, we present our new word detection algorithm that combines user behaviors in section 4. Experiment results and discussions are showed in section 5. Finally, section 6 concludes the whole paper and gives some future works.

2 Related Work

2.1 New Word Detection

Generally speaking, new word detection has been considered in two fields of research: first, new word detection aims to find the words that have not emerged before, which is similar to the definition of out-of-vocabulary words. Second, new word detection is always performed simultaneously with Chinese word segmentation task to achieve better results. Many techniques with respect to the problem of Chinese new word detection have been proposed in the past, which roughly can be grouped into two classes: rule-based methods and statistical machine learning methods.

As for rule-based methods, [Wang *et al.*, 1995] used an n-grams based approach to identify and classify Chinese unknown words. [Chen and Bai, 1998] took new word detection and segmentation as a whole process and proposed a rule-based system which has been proven to be powerful.

Along with the development of statistical learning, more and more techniques have been applied in the new word detection task. [Li, *et al.*, 2004] defined it as a binary classification problem and applied an SVM classifier to identify Chinese new word. In their work, many features, including in-word probability of a character, anti-word list, frequency in documents, etc., have been constructed to solve this problem. [Peng, *et al.*, 2004] brought conditional random filed into Chinese new word identification and further improved the quality of word segmentation by adding the newly detected words into the vocabulary. [Zhang, *et al.*, 2002] defined a set of word roles to detect unknown words in real texts based on role tagging method. [Kleinberg, 2002] focused on modeling word bursts in the continuous stream data using a probabilistic automaton.

Different from the methods introduced above, we proposed a novel method which mainly considers the user behaviors instead of Chinese word segmentation or feature representation of words. In fact, CWS itself is a very tough and challenging problem. On the other hand, our method follows a straight collaborative way, which is easy to understand and achieves acceptable performances. Moreover, our method follows an unsupervised strategy.

2.2 Collaborative Filtering

Collaborative filtering [Deshpande and Karypis, 2004] is the procedure of filtering for useful information or items using various machine learning techniques. It has been applied to many different kinds of applications such as recommendation of movies, music and books. Amazon and Netflix are two typical commercial sites that implement collaborative filtering systems. These systems try to predict the interests of users according to their historical taste, and then recommend useful items, such as books or films to them. According to [Breese *et al.*, 1998], collaborative filtering algorithms can be mainly classified into memory-based and model-based classes. Memory-based methods [Breese *et al.*, 1998; Nakamura and Abe, 1998] make predictions using the previous rated items by the users. Model-based methods [Breese *et al.*, 1998; Hofmann, 2003], on the other hand, build a model according to the collection of ratings, then make rating predictions by the learned model.

The underlying assumption of collaborative filtering is that users who have the same taste in the past tend to agree in the future. Usually, the system searches similar users based on the historical item taste. We can regard it as item to user step. The next step is to recommend items liked by most of the similar users. In fact, this is a user to item step. We use the similar manners in this paper. First, we find the potential experts according to how many representative words they used. This can be recognized as a word to user step. Then we detect the new words that used by most of those experts. This is a user to new word step. It can be seen clearly that our method follows the collaborative filtering methodology.

2.3 Google Sets and Bayesian Sets

Google Sets [Google] is a remarkably interesting tool that takes several words as input, and then it can retrieve other related words as output. Generally speaking, it works well with input words in English language. But the performance decreases rapidly in other languages such as Chinese. The experiment in section 5 proves this.

Bayesian Sets [Ghahramani and Heller, 2005] addressed the same related word retrieval task in the framework of Bayesian inference. Bayesian Sets computes a score for each candidate word by comparing the posterior probability of that word given the input words, to the prior probability of that candidate word. We can easily extend our method to the same related word retrieval task by taking user behaviors into account. Experiment results show that our method and Bayesian Sets gives better performance.

3 Background

In this section, we will first introduce some background knowledge, including the experiment datasets, user dictionary, and cell dictionary.

All the resource used in this paper is generated from Sogou Chinese pinyin input method [Sogou, 2006]. We use Sogou for abbreviation hereafter. Sogou introduced a novel concept named **cell dictionary**. Cell dictionary includes terminologies in a certain field. Like the mode of Wikipedia, everyone can create or modify a cell dictionary. Cell dictionary is somehow similar to domain-specific lexicon. As domain-specific lexicon is always created by domain experts, it is quite costly to obtain and maintain.

Users can gain better experience by choosing appropriate cell dictionaries. For example, experts in computer science can choose computer cell dictionary to help them type terminologies in computer science faster and more accurately.

On the other hand, **user dictionary** records the word lists that users have used on their computers. Volunteers are also encouraged to upload their anonymous user dictionaries to the server side. Furthermore, volunteers can also upload configuration information, such as cell dictionaries they chose. In order to preserve user privacy, usernames are hidden using MD5 hash algorithm. So, the dataset we used is completely anonymous and user privacy is safeguarded.

4 New Word Detection Algorithm

As discussed before, starting from the cell dictionaries, we try to find the potential experts in certain fields through their configuration information and user dictionaries. On the other hand, cell dictionaries always contain some noise, such as words not related to particular field, or some missing words that should be included. In fact, we attempt to identify those missing words by investigating the community of potential experts discovered in the first step. The basic idea is that words used much more frequently in this community than others are most probably new words.

In the next subsections, we will first introduce how to select most representative words from cell dictionaries. Then we discover potential experts who use those words quite often. Finally, we detect new words from behaviors of those experts. In brief, our proposed method performs in a collaborative filtering way, from words to find professional users, then from users to detect new words. While in recommendation systems using collaborative filtering techniques, we are always under the similar circumstances. Based on the historical items that a user has picked, which item should be recommended? Usually, we first try to find similar users that have the same item taste, which is an item to user step. Then, only the items that are mostly liked by those similar users would be recommended, which is a user to item step.

4.1 Representative Word Selection

Cell dictionaries are maintained by everyone, which brings inevitable noise. For example, words should not be included, while some missing words that should be there are not included. In this subsection, we focus on how to select representative words in cell dictionaries, and those words will be useful for potential experts searching in the next step. We use positive instances to represent the users who have used certain cell dictionary, while negative instances represent the other users. We label users as positive or negative instances according to their configuration information and the corresponding cell dictionary. The same user can be labeled as positive or negative based on different cell dictionaries.

Inspired by [Li and Sun, 2007], we consider two factors of words, which are discriminability and coverage, respectively. Coverage refers to the popularity of a word, or how many users have used this word. We can obtain this information from user dictionaries. On the other side, discriminability means how unbalanced the distribution of the word in positive and negative users is. A representative word should have powerful discriminability which can distinguish positive users from negative ones. Meanwhile, it should maintain wide enough coverage. We will discuss those two factors separately in the following part.

[Forman, 2003] proposed a straightforward metric named probability ratio to measure the discriminability of a word: (for brief, we use PR instead hereafter)

$$PR(w,c) = \frac{P(w|c_+)}{P(w|c_-)} = \frac{uf(w,c_+)/uf(c_+)}{uf(w,c_-)/uf(c_-)} \qquad (1)$$

As shown above, $uf(c_+)$ refers to the number of positive users, while $uf(w,c_+)$ refers to the number of positive users who have used word w. $uf(w,c_-)$ indicates how many negative users that have used w, and $uf(c_-)$ is the number of negative users.

For coverage, we simply use number of users who have used word *w* to measure the coverage of *w*. In this paper, we use $uf(w)$ to represent coverage of *w*. Generally speaking, discriminability and coverage have a slightly negative correlation. Widely used words always have poor discriminability. It is not reasonable to use those words to separate positive users from negative ones, because almost everyone has used them. On the contrary, a word that mainly appears in positive users shows strong discriminability. However, this particular word has weak coverage with few occurrences among negative users. In this paper, we take both discriminability and coverage into consideration. A parametric representative word selection criterion is defined as follows:

$$\zeta(w,\lambda) = \left(\frac{\lambda}{\log(PR(w,c))} + \frac{1-\lambda}{\log(uf(w))} \right)^{-1} \qquad (2)$$

A weight harmonic average formulation is adopted here to balance the two factors. $\lambda \in [0, 1]$ is the weight for log(PR) which indicates the importance of discriminability. In this paper, we treat discriminability and coverage fairly. In other words, we set λ to 0.5. In the following experiments, we will show top ranked words according to the three selection criterions respectively.

4.2 Potential Expert Search

For each cell dictionary, we can select the most representative words according to the strategy discussed above. To put it more specifically, words in cell dictionary are ordered according to their scores obtained from formulation 2. Then top-*d* words are picked up. In this paper, we select top-1000 words as most representative words.

On the other hand, it is reasonable to infer that users who use the corresponding representative words very frequently are potential experts in the certain field. For instance, if 90 percent of those words emerge in one user's dictionary, we are confident enough to believe that this user is a potential expert. In fact, we sort those potential experts according to the percentages of top-*d* words they have used. We never consider all the words in cell dictionary, because it is noisy. For example, cell dictionary on computer science contains word such as "下载" (download) and "问题" (problem), which are noisy words. It is unreasonable to label a user as a computer science expert if he uses these two words very frequently. On the other hand, words like "人工智能" (artificial intelligence) and "机器学习" (machine learning) usually only appear in computer science community. If words like those emerge in user dictionary, he is very likely to be an expert in computer science.

4.3 New Word Detection

As mentioned before, cell dictionaries are maintained by humans. Inevitably, some words that should be included are missing. We regard those missing words as new words in this paper, and we are trying to identify them.

In the previous step, we have found the potential experts in certain field. The next step is to detect new words from their behaviors. Suppose a word *w* is widely used among the community of those experts, for example, *w* appears in half of those experts' user dictionary. Meanwhile, *w* seldom occurs in other users' dictionary. We can fully believe that *w* is a missing terminology that should be included.

In brief, we detect new words according to their ratio of distributions between potential experts and other users. The more frequently it appears in potential experts, the more likely it is a missing new word. To give a brief demonstration, we will show some examples in the experiment section.

As a matter of fact, we can follow the similar procedure described before, but start from only one discriminative word, find the corresponding users who used this very word, and then mark them as potential experts. Finally, we can find words used significantly more frequently in those experts than others, which are related words of that particular word. Detailed experiment results are shown in section 5.

5 Experiments

5.1 Experiment Setting

The datasets we used in the following experiments mainly include user dictionaries as well as cell dictionaries. User dictionaries record the words that uses have used. Cell dictionaries contain all kinds of terminologies in various fields. To give a brief and convincing example, we only take a cell dictionary on computer science as a sample.

First, we will show the results of representative words selection. It is considered to be more reasonable by taking both coverage and discriminability into account. Then, we selected top-1000 words as the most representative set. Furthermore, we search the potential experts on computer science based on their user dictionaries and the representative words identified in the previous step. Finally, words used much more frequently in those experts are identified as new words in computer science. We will show some selected new words in the next subsection. Moreover, we also extend our algorithm to single discriminative word case. The goal is to discover related words. We found some interesting results which indicate that our method is effective.

5.2 Representative Word Selection

In this subsection, we will show some results about representative words selection. As discussed in subsection 4.1, we mainly focus on two characteristics of words: coverage and discriminability. Coverage reflects the popularity of particular word, or the number of users who use this word. While discriminability measures how we can distinguish corresponding users from positive and negative. We will show some examples with good coverage or discriminability respectively. Finally, better results can be gained by considering those two characteristics.

Table 1 shows some most widely used words in computer science cell dictionary. The most popular word is "问题" (problem), which makes up nearly 80 percent of the whole user group. Some of them are noisy which have nothing to do with computer science, such as "联系" (contact). However, this is inevitable because cell dictionaries are maintained by human, and everyone can do modifications on them, which will introduce some noise. On the other hand, although the rest words are related to computer science to certain degree, they have very poor discriminability, because almost everyone uses them. It is not reasonable to label a user as computer science expert if he uses those words very frequently.

Rank	Word	User Coverage
1	问题(problem)	78.33%
2	下载(download)	73.94%
3	照片(picture)	67.25%
4	空间(space)	66.73%
5	联系(contact)	66.53%

Table 1: Top words Ordered by Coverage

Table 2 shows some most discriminative words in computer science cell dictionary. Those words are very professional terminologies. Unlike the widely used words shown in table 1, once a user typed these words, it is very likely that he is an expert in computer science. But the shortcoming of them is that they are rarely used. For example, only about 3 users use word "并行数据库" (parallel database). That is to say, if we use these words as representative members, we can only discover very rare potential experts. Furthermore, this has a very bad impact on the next new word detection step.

Rank	Word
1	并行数据库(parallel database)
2	上下文无关语言(context free language)
3	个人词语(personal word)
4	信息浏览服务(browsing service)
5	假通过率(pseudo pass rate)

Table 2: Top words Ordered by Discriminability

As stated before, we should consider both coverage and discriminability in representative words selection step. In this paper, we equate the two factors. This indicates that λ is set to 0.5 in formula 2. Table 3 shows the corresponding results.

Words listed in Table 3 have powerful discriminability as well as wide enough coverage. On one hand, they are more popular than words listed in table 2. In fact, they are more general terminologies in computer science. On the other hand, they maintain better distinguishing ability than words listed in table 1. We can be confident to search enough number of potential experts using those words. In this paper, we select top-1000 words like those as the most representative set.

Rank	Word
1	重定向(redirect)
2	转义(escape)
3	易用性(accessibility)
4	应用层(allocation layer)
5	可扩展性(scalability)

Table 3: Top words Ordered by Coverage and Discriminability

5.3 Detecting New Word

In the next step, we will search the potential experts using the top-1000 representative words. The more representative words a user uses, the more likely he is a potential expert. We select 8000 potential experts in this experiment, and then detect new words by analyzing the user behaviors in the community of those experts.

Generally speaking, experts in the same professional field may use the same words to a certain extent, while other users seldom use them. For this consideration, we detect new words based on how unbalanced the distributions of them in the community of potential experts and other users are. We also take the popularities of candidate words into account. Table 4 shows some new words detected by our method.

跨平台(cross platform)	基类(base class)
复杂度(complexity)	实例化(instantiation)
类库(class library)	缓冲区(buffer)
头文件(header file)	递归(recursion)
端口号(port number)	服务器端(server side)
死循环(infinite loops)	子目录(subdirectory)

Table 4: Detected New Words

As seen in the above table, we indeed detect some new words about computer science which are not included in original cell dictionary. We also observe that those newly detected words are widespread among specialists on computer science. For example, "死循环" (infinite loops) and "基类" (base class) are widely used by programmers. "端口号" (port number) may be used by network engineers.

We report the performance of our method using different evaluation metrics in Table 5. We detect 1,629 new words in total. Five persons who have rich background knowledge in computer science are asked to judge whether a word is related to computer science or not. The final results are made based on their votes.

Because we are not able to get a complete list of computer science related words, we use the binary preference measure [Buckley and Voorhees, 2004] to evaluate our method with incomplete information. For a topic with judged R relevant documents where r is a relevant document and n is a nonrelevant document, Bpref is defined as follows:

$$Bpref = \frac{1}{R}\sum_{r} 1 - \frac{|\,n \text{ ranked higher than } r\,|}{R} \quad (3)$$

Results indicate that our method gains up to 0.86 in accuracy. Moreover, it is quite promising that our method shows 1.00 and 0.92 in P@30 and P@100 respectively.

Bpref	Accuracy	P@10	P@30	P@100	P@1000
0.56	0.86	1.00	1.00	0.92	0.88

Table 5: Performance on Computer Science Related Dataset

Based on the above analysis, we are now confident to claim that our method achieves an acceptable performance and it is quite helpful to incorporate user behaviors in new word detection.

5.4 Related Word Retrieval

In this subsection, we will show some experiment results about detecting related words by starting from single discriminative seed word. Detailed steps have been discussed in subsection 4.3. Here we give some brief explanations as well as some interesting results.

Table 6 shows some selected words related to "人工智能" (artificial intelligence). On the whole, all these words are relevant to artificial intelligence. Words like "遗传算法" (generic algorithm) and "回溯" (backtracking) are basic concepts in artificial intelligence field. Some words are related to certain research areas, such as "机器学习" (machine learning) and "数据挖掘" (data mining). To sum up, our method gains reasonable results in related words detection.

机器学习(machine learning)	分类器(classifier)
模式识别(pattern recognition)	算法导论(introduction to algorithm)
神经网络(neural network)	信息检索(Information Retrieval)
图灵(Turing)	回溯(backtracking)
数据挖掘(data mining)	面向对象(object oriented)
遗传算法(generic algorithm)	形式化(formulization)

Table 6: Words Related to "Artificial Intelligence"

To further investigate the performance of our method, we carry out comparisons with Google Sets and Bayesian Sets. We randomly select 100 seed words that related to computer science. Top 10 results of each method are used to do evaluation. We measure the mean reciprocal rank of the first retrieved result (MRR), Bpref as well as precision at 5 and 10 in Table 7. For a sample of queries Q, rank_i is the rank of the first relevant result for query Q_i, MRR is defined as follows:

$$MRR = \frac{1}{|Q|}\sum_{i}\frac{1}{rank_i} \quad (4)$$

	Bpref	MRR	P@5	P@10
Google Sets	0.43	0.65	0.44	0.35
Bayesian Sets	**0.47**	0.70	**0.51**	0.44
Our Method	0.46	**0.71**	0.49	**0.45**

Table 7: Comparisons with Google Sets and Bayesian Sets

As shown in Table 7, we can clearly see that our method gains quite comparable results to Bayesian Sets. We also find that Google Sets give relatively poor results with words in Chinese language. In fact, Google Sets does not return any results for words "加权" (weighting), "框图" (block diagram) and "分词" (word segmentation) in our experiment.

6 Conclusion

In this paper, we proposed a novel method to detect new words by incorporating user behaviors. Unlike traditional new word detection methods, our method follows a collaborative filtering strategy: first we start from a single discriminative word or a group of selected representative words, and then we search for the potential experts on certain fields. Finally, we focus on the user behaviors of those experts. We

believe that words used much more frequently in the community of experts than others are very likely to be new words. Experiments on computer science filed indicate that our method is effective and gains up to 0.86 in accuracy. Moreover, we obtain some interesting new words by starting from a single word. Our method gives comparable performance to Bayesian Sets and better performance than Google Sets. However, we also observed some noisy new words in the experiment. We plan to use some background knowledge such as Part-of-Speech information to reduce those wrongly detected new words in the future. Furthermore, the method is language independent and can be extended to other languages once we have user dictionaries in the target language.

Acknowledgement

We thank the following people for helpful comments on previous drafts: Yan Zhang, Xiance Si, Qixia Jiang, Shaohua Teng and Kaixu Zhang. This work is supported by a Tsinghua-Sogou joint research project and the National Science Foundation of China under Grant No. 60621062 and 60873174.

References

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

[Breese *et al.*, 1998] John S.Breese, David Heckerman and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*

[Buckley and Voorhees, 2004] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25-32

[Chen and Bai, 1998] Keh-Jiann Chen and Ming-Hong Bai. Unknown word detection for Chinese by a corpus-based learning method. *International Journal of Computational linguistics and Chinese Language Processing*, 3(1): 27-44

[Deshpande and Karypis, 2004] Mukund Deshpande and Geogre Karypis. Item-Based Top-N Recommendation Algorithms, *ACM Trans. Information Systems*, 22(1), 143-177

[Forman, 2003] George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3(1): 1289-1305

[Ghahramani and Heller, 2005] Zoubin Ghahramani and Katherine A. Heller. Bayesian Sets. In *Advances in Neural Information Processing Systems*

[Google] Google. Google Sets. Accessed on Jan. 10th, 2009, available at: http://labs.google.com/sets

[Hofmann, 2003] Thomas Hofmann. Collaborative filtering via Gaussian probabilistic latent semantic analysis, In *Proceedings of the 26th Annual International Conference on Research and Development in Information Retrieval*, pages 259-266

[Kleinberg, 2002] Jon Kleinberg. Bursty and Hierarchical Structure in Streams, In *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91-101

[Li and Sun, 2007] Jingyang Li and Maosong Sun. Scalable term selection for text categorization, In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 774-782

[Li, *et al.*, 2004] Hongqiao Li, Chang-Ning Huang, Jianfeng Gao, and Xiaozhong Fan. The use of SVM for Chinese new word identification. In *Proceedings of the First International Joint Conference on Natural Language Processing*, pages 497-504

[Linden *et al.*, 2003] Greg Linden, Brent Smith, and Jeremy York. Amazon.com Recommendations: Item-to-Item collaborative filtering. *IEEE Internet Computing*, 7(1): 76-80

[Miller *et al.*, 2003] Bradley N. Miller, Istvan Albert, Shyong K. Lam, Joseph A. Konstan, and John Riedl. MovieLens unplugged: experiences with an occasionally connected recommender system. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pages 263-266

[Nakamura and Abe, 1998] Atsuyoshi Nakamura and Naoki Abe. Collaborative filtering using weighted majority prediction algorithms. In *Proceedings of 15th International Conference on Machine Learning*, pages 395-403

[Peng, *et al.*, 2004] Fuchun Peng, Fangfang Feng, and Andrew McCallum. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 562-568

[Sogou, 2006] Sogou, Sogou Chinese Pinyin Input Method. available at http://pinyin.sogou.com/

[Sproat and Emerson, 2003] Richard Sproat and Thomas Emerson, First International Chinese Word Segmentation Bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*

[Wang *et al.*, 1995] Mei-Chu Wang, Chu-Ren Huang, and Keh-Jiann Chen. The identification and classification of unknown words in Chinese: an n-grams-based approach. *Festschrift for Professor Akira Ikeya*, pages 113-123, Tokyo: The Logico-linguistics Society of Japan

[Zhang, *et al.*, 2002] Huaping Zhang, Qun Liu, Hao Zhang, and Xueqi Cheng. Automatic recognition of Chinese unknown words based on roles tagging. In *Proceedings of The First SIGHAN Workshop on Chinese Language Processing*, pages 71-77