New Approach for Automated Categorizing and Finding Similarities in Online Persian News

Çevrimiçi Farsça Haberlerde Otomatik Sınıflandırma ve Benzerliklerin Bulunması için Yeni Bir Yaklaşım

Naser Ezzati Jivan

Managing Director, National Library and Archives of the I.R of Iran, Ezzati@nlai.ir

Mahlagha Fazeli

Computer Engineer, Iranian University of Science and Technology, Mfazeli@comp.iust.ac.ir

Khadije Sadat Yousefi

Computer Engineer, Iranian University of Science and Technology, Khyousefy@comp.iust.ac.ir

Abstract: The Web is a great source of information where data are stored in different formats, e.g., web-pages, archive files and images. Algorithms and tools which automatically categorize web-pages have wide applications in real-life situations. A web-site which collects news from different sources can be an example of such situations. In this paper, an algorithm for categorizing news is proposed. The proposed approach is specialized to work with documents (news) written in the Persian language but it can be easily generalized to work with documents in other languages, too. There is no standard test-bench or measure to evaluate the performance of this kind of algorithms as the amount of similarity between two documents (news) is not well-defined. To test the performance of the proposed algorithm, we implemented a web-site which uses the proposed approach to find similar news. Some of the similar news items found by the algorithm have been reported.

Keywords: Categorization of web pages, automatic categorization of Persian News, feature, similarity, clustering, structure of web pages

Öz: Web, web sayfaları, arşiv dosyaları ve görüntüler gibi verilerin farklı formatlarda dosyalandığı büyük bir bilgi kaynağıdır. Web sayfalarını otomatik olarak kategorize eden algoritma ve araçların gerçek hayatta yaygın uygulamaları vardır. Farklı kaynaklardan haber toplayan bir web sitesi bu tür durumlara bir örnek olarak verilebilir. Bu bildiride haberleri kategorize etmek için bir algoritma önerilmektedir. Önerilen yaklaşım Farsça yazılmış belgelerle (haberler) çalışacak şekilde özelleştirilmiştir ancak, diğer dillerde yazılmış belgelerle de çalışacak şekilde kolayca genelleştirilebilir. İki belge (haber) arasındaki benzerlik düzeyi iyi tanımlanmadığından bu tür algoritmaların performansını ölçecek ya da test edecek standart bir test ya da ölçü yoktur. Önerilen algoritmanın performansını test etmek için önerilen yaklaşımı kullanarak benzer haberleri bulmak için bir web sitesi kurduk. Algoritma tarafından bulunan benzer haber maddelerinden bazıları bildiride rapor edilmektedir.

Anahtar sözcükler: Web sayfalarının kategorizayonu, Farsça haberlerin otomatik kategorizayonu, özellik, benzerlik, kümeleme, web sayfalarının yapısı

Introduction

Taking into account the large bulk and wide variety of web data, organizing these data for easy access and improving the search results is vital. Many efforts have been made to categorize web pages. As the web includes many different kinds of data (such as texts, images, multimedia data, etc.), there are different categorization methods for each.

These methods include: categorization of texts based on statistical and algorithmic methods of machine learning (Kwon & Lee, 2000). In machine learning algorithms, training data are used to train categorizers. The software used for categorizing news is called categorizer. Categorizers can categorize new pages after they have been trained. These

methods include k-Nearest Neighbor approach (Yang & Lui, 1999), Bayesian probability models (Lewis & Ringuette, 1994; McCallum & Nigam 1998; Combarro et. al., 2005), inductive learning rules (Apte, Damerau, & Weiss 1994), backup machines (Dumais, Platt, Heckerman, & Sahami, 1994), neural networks (Weigend, Weiner, & Peterson, 1999), and decision making trees (Lewis & Ringuette, 1994).

The possibility of reading online news is one of the web facilities used by many users. There are a lot of sites which include daily news. If a user wants to read more about a piece of news from other sites, he or she will have to search different sites to find similar news. This can be time-consuming for the user. To solve the problem, some methods have been proposed for categorization of news on the internet.

In this project, news headlines and summaries were used for categorization of news. Keywords for each piece of news are extracted from the headline and summaries, which are then used to collect similar news from a news data bank using a web crawler. The first part of this paper examines methods of web page categorization and in the second part approach used in the paper for categorizing Persian news is elaborated on. In the third part, the results of trying the approach out for Persian news are presented.

Categorization of Web Pages

Many efforts with differing degrees of precision have been made for categorization of web pages, the most important of which include:

- Manual categorizations by experts
- Cluster methods
- Content analysis of links and documents

Manual Categorization

In the first method, some experts in each field analyze the contents of web pages and put them into different categories according to their topics. A good example of this method of categorization is that of dmoz.org which categorizes web pages using experts around the world. Yahoo had used the method before 1998 (Kwon & Lee, 2000). Although this method has a high degree of precision, the increasing number of web pages entails using a larger number of experts, making it very difficult and impractical.

Cluster Methods

Clustering pages is used for automatic categorization of web pages. Each document is a web page and each cluster includes many documents. The first phase of clustering is extracting features. For each document the general words are first omitted. These are the words without independent separate meaning, like prepositions. Each feature is a keyword or phrase appearing in a group of documents. Keywords of a document can be extracted using different approaches like that of tf (Guha, Rastogi, & Shim, 1998). Next, each document is shown using a Feature Vector, which includes the features of the document and the numerical value of each feature. The numerical value shows the frequency of the feature in the document. After the formation of the feature vector, the clustering algorithm is used on the collection of vectors to categorize the documents. Examples of clustering algorithms include: BIRCH (Tokunaga & Makoto, 1994),CURE (Zhang, Ramakrishnan, & Livny, 1996) DCTree (Guha et al., 1998).

Structural Categorization of Web Pages

The structure of web pages is used widely to improve the organization, search and analysis of information on the web. As an example a link is intended presumably to show the topical interrelations between two documents. On this basis, the texts of web pages are divided into three groups:

- The anchor text used for description of the link
- The text near the link which usually includes around 25 words before and after the link (Chan, Sun, & Lim, 2001)
- The regular text that forms the remainder of the page

In most web pages the anchor text and the text around the link are better descriptors of the topic of the page. Common methods of web page categorization which use word or phrases of the destination page pay attention to keywords of the

page as well as the anchor text and its surrounding text to extract features. For example, Google includes the pages that have the searched keywords in texts around their links to improve search results even if the rest of the page does not have the words.

Automatic Categorization of News Using S-V-M Categorizer

The architecture of this approach includes six main modules: pre-processing, presentation, storing, S-V-M, user registration, and retrieval of web pages.

A Web page retrieval module downloads online articles and news using crawling robots from news sites. The pre-processing module includes text separators, pre-processor of documents, and generator of text vectors. The pre-processor of document omits general words and gives the remainder to the generator of the vector to make the vectors using tf*idf (Chan et al., 2001). The final product goes to the S-V-M module. Each text vector includes the remaining words of the text. There are three news databases, Reuter's tests, and the system in the storing module. The news database stores the features of the news such as date address and text of the news, downloaded from news websites. The system database stores information about users and groups related to each person. The Reuter's test battery is used to train the S-V-M web page categorizer. The next module is S-V-M.

This module is a binary categorizer including an S-V-M trainer and S-V-M categorizer to train S-V-M; a category (like sport) is selected and its related model file is produced. The model file is delivered to the S-V-M categorizer which performs categorization of the downloaded documents. The representation model shows the categorization results based on the hierarchy or priorities specified by the categorizer. The registration module manages users' information, and their personal groups.

The categorizer software performs categorization using two methods, general and specific, which are described below.
1. In the general categorization, all training documents are chosen from Reuter's documents. Ten general groups are now supported by the categorizer. An S-V-M categorizer has been produced for each group. After training, the output of the S-V-M categorizer is saved in the system database.
2. The news articles are downloaded from their sources and their extracted texts are stored in the news database.
3. When the user requests news of a special group, the recently downloaded documents are retrieved from the news database and the vectors of each document are created by the pre-processor module.
4. An S-V-M categorizer is created for each group and their model files are stored in the system database.
5. When a user requests a piece of news in his personal group, the recently downloaded pieces of news are retrieved and their document vectors are created.
6. The document vectors and the group model file are given to the S-V-M of the group and the results of the categorization are shown in order of priority.

The system is designed in such a way that when a user reads a piece of news from a specific group, he can click on the "related" button if it is the related news he wants. This feedback will be used for further training to improve the categorization (Chan et al., 2001).

Automatic Categorization of Persian News

In this part, the implementation of the project for categorization of Persian news is explained. This is done based on the features of Persian syntax and includes two phases. In the first phase, the features are extracted and stored in the local database. In the second phase, the similar pieces of news are extracted using the specified features. In Figure 1, the structure of the automatic categorizer of Persian News is shown.

Extraction of Features

Persian sentences include general and keywords based on their semantic load. Keywords carry the general meaning of the sentence and general words are used with keywords to complement the meaning of the sentence. In this approach, we specify keywords, the topic of the news, and its date as the features of the Persian sentences. Therefore, those pieces of news which have similar keywords and topics are regarded as related and similar.
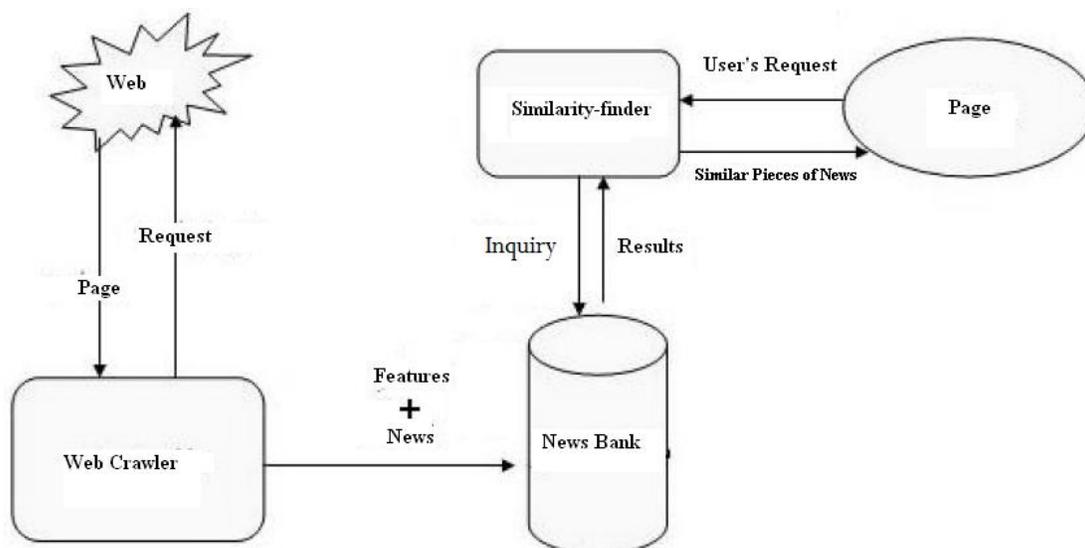
Figure 1. Architecture of Persian news categorizer

Omitting General Words:

General words are divided into two groups based on how they can be recognized in the sentence:

1. Words which are recognizable using the sentence structure.
2. Words which are specified using a list of general words

In order to find the general words based on the sentence structure, Persian sentence grammar was used as shown in Figure 2.

```
<title>  → <sentence> (<dot>)
<sentence>→ <word> { <space> <word> <space>}
<word>→ <general word> |<key word>
<key word>→ <letter><letter><ketter>{<letter>}
<general word>→ <verb>| <mark> |<additional word>
<additional word>→ <two word>|<other additonal word>
<two word>→ <letter> <letter>
<other additional word >→"همه"|"برای"|"زیرا"|"این"|"چرا"|"های"|...
<verb>→(<past mark>|<present mark>|<future mark>)<normal verb>
<normal verb>→<keyword>
<past mark>→"شده"(بودند|بودیم"|"بودند"|"بودی"|"بودم")
<present mark>→"می"|"نمی"
<dot>→"."
<space>→" "
<letter>→ "الف"|...|"ی"
<mark> → ":"|"?"|";"|","
```

Figure 2. Persian sentence grammar for extracting features

The following rules can be deduced based on this grammar:

1. The word before the period in each sentence
2. The word after a conjugation of the future tense of some verbs using the auxiliary verb "خواستن": خواهم رفت، خواهم دید

The word coming before the auxiliary verb "بودن" in the past tense, like: رفته بودم، دیده بودي

The words which are created from at most two letters are general words and are not worth categorizing and thus are omitted.

Other groups of general words include those that are used in most sentences or are not semantically worthy, so they are omitted from the news.

Omitting general words can be done in two ways:

1. Offline: after the crawler gets the news from the related site
2. Online: when the crawler requests similar pieces of news

In the first method when the crawler gets the news from the site, its general words are removed and the remaining words that are features of the news are stored in the database. In the second method, just the news is stored in the news bank and upon the user's request its general words are omitted. Then, based on the keywords, similar pieces of news are specified in the news bank. The first approach needs more memory because features of all pieces of news must be stored. But the similar pieces of news are found with higher speed. In the second method, lower memory is needed but the speed of finding similar news is reduced. As the higher speed is more significant than the memory needed, in this project, the first method was used for implementation.

Finding Similar Pieces of News

There are different definitions for similarity between the pieces of news. The similarity can be semantic or syntactic. Similarity of news is a relative concept and can be different from different points of view. For example, for some people, those pieces of news which have the same intention or content are similar, while others may consider a common word as the reason for similarity of different pieces of news.

For the project to be comprehensive, those pieces of news which have at least one similar word are considered similar. As stated before, the features of each piece of news are stored with the news in the news table. Therefore, by creating inquiries which look for pieces of news similar to the one at hand, and performing them on the news table, similar pieces of news are gathered. The inquiries are performed in order of priority on the news table and the results are shown to the user.

The priorities of inquiries for a piece of news with n features include:
The first priority: inquiries with n features
The second priority: all inquiries with permutations of n-1 features
The third priority: all inquiries with permutations of n-2 features

…
The nth priority: all inquiries including only one of the keyword.

Assessment

The similarity-finder software has been implemented in two parts, crawler and similarity-finder. The crawler refers to sources of news in the internet at predefined points of time and it downloads different pieces of news based on their topics and dates. In this part, the features of each news items with regard to its summary (here, the first paragraph of the news) topic, and the download time are stored in the site's local database.

In the second part, the similarity-finder, the news items similar to the particular one are searched and retrieved. This part, first, extracts the features of the selected piece of news, based on which it creates appropriate inquiries and looks for similar news in the news database. These inquiries are created based on the priorities and permutations of features of the news.

The crawler system and similarity-finder are implemented in Linux operating system using Php language, some of whose results are represented in Figure 3 below.

اخبار مشابه برای خبر : | وزیر صنایع و معادن ایران عازم کشور سوریه شد |

وزیر صنایع و معادن ایران عازم کشور سوریه شد - خبرگزاری جمهوری اسلامی ایران – ۲۰ ساعت و ۵۰ دقیقه قبل

وزیر صنایع و معادن ایران عازم کشور سوریه شد - خبرگزاری دانشجویان ایران – ۲۲ ساعت و ۲۵ دقیقه قبل

معاون وزیر صنایع و معادن به عنوان مدیرعامل جدید شرکت آلومینیوم ایران معرفی شد - خبرگزاری دانشجویان ایران -۲ روز و ۶ ساعت قبل

وزیر صنایع و معادن خواستار شد :انتقال دانش طراحی بدنه خودرو از سوی رنو به سایپا و ایران‌خودرو - خبرگزاری دانشجویان ایران – ۳ ماه و ۷ روز قبل

وزیر صنایع و معادن به نروژ رفت - خبرگزاری دانشجویان ایران – ۳ ماه و ۷ روز قبل

Similar pieces of news for: The Iranian Minister of industries and mines left for Syria

The Iranian Minister of industries and mines left for Syria-Islamic Republic News Agency – 20 hours and 50 minutes ago

The Iranian Minister of industries and mines left for Syria – Iranian Students News Agency- 22 hours and 25 minutes ago

Deputy Minister of industries and mines was introduced as the new manager of Iranian Aluminum Company – Iranian Students News – 2 days and 6 hours ago

Minister of industries and mines suggested: transfer of knowledge of designing car body from Renault to SAIPA and Iran Khodro- Iranian Students News – 3 months and 7 days ago

Figure 3. Results of similarity-finder software for Persian news (and English translation)

Evaluating the results obtained by the software and their manual examination showed that using the permutations in headline, summary and topic of news has a precision of 79 percent. This figure is the result of manual examination of 100 pieces of news and similar news items and comparing them with the results obtained by the similarity-finder software for these pieces of news.

Summary and Future Work

In this article the implementation of similarity-finder software for Persian news is elaborated on. First, different methods for categorization of web pages and web news are examined. Then the method implemented for Persian news is elaborated on. Finally, the results of implementing the software and its manual examination are presented.

Since in this system only the resemblance in subject and keywords in the news have been used, the results are not 100 percent accurate. A method of solving this problem to obtain more accuracy in the search results is to use semantic similarities to find similar pieces of news. For this to happen, the features must be chosen for both keywords and concepts of the news. The semantic features may have little in common with keywords and just take into account the meaning of a piece of news. For example, if there is the word "برانکو" (the former coach of Iran's football team) in the news text, the pieces of news about Iran's football team must be extracted as well. But, based on keywords, only those pieces of news which have similar words and general topics are presented as similar without regard to semantic similarities.

With regard to the results of evaluating the similarity-finder software based on keywords it is expected that using semantic similarities helps us obtain a high precision-above 79%- in finding similar pieces of news and categorizing them. Implementation of a semantic similarity-finder can be the next generation of categorizers of Persian news as compared with the similarity finders based on keywords.

References

Apte, C., Damerau, F., & Weiss, S.M. (1994). Automated learning of decision rules for text categorization, *ACM Transactions on Information Systems*, *12*(3), 233-251.

Chan, C.-H., Sun, A., & Lim, E.-P. (2001). Automated online news classification with personalization, Center for Advanced Information Systems, Nanyang Technological University Nanyang Avenue, Singapore, 639798. Retrieved August 02, 2010 from http://ncsi-net.ncsi.iisc.ernet.in/gsdl/collect/icco/index/assoc/HASH01de.dir/doc.pdf

Combarro, E.F., Montanes, E., Diaz, I., Ranilla, J., & Mones, R. (2005). Introducing a family of linear measures for feature selection in text categorization. *IEEE Transactions on Knowledge and Data Engineering*, *17*(9), 1223-1232.

Dumais, S.T., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization, In *Proceedings of the Seventh International Conference on Information and Knowledge Management (CIKM'98)*, (pp.148-155). New York: ACM.

Guha, S., Rastogi, R., & Shim, K. (1998). Cure: an efficient clustering algorithm for large database, In *Proceedings of the ACM SIGMOD Conference on Management of Data* (pp. 73-84). Seattle: ACM.

Kwon, O.-W. & Lee, J.-H. (2000). Web page classification based on k-nearest neighbor approach. In *Proceedings of the fifth international workshop on information retrieval with Asian languages* (pp. 9-15). New York: ACM.

Lewis, D.D. & Ringuette, M. (1994). A classification of two learning algorithms for text categorization, In *Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*, (pp.81-93). Las Vegas, USA.

McCallum, A. & Nigam, K. (1998). A comparison of event models for Naïve Bayes text classification, In *AAAI-98 Workshop on Learning for Text Categorization*.

Tokunaga, T. & Makoto, I. (1994). Text categorization based on weighted inverse document frequency, Special Interest Groups and Information Process Society of Japan (SIG-IPSJ).

Weigend, A.S., Weiner, E.D. & Peterson, J.O. (1999). Exploiting hierarchy on text categorization, *Information Retrieval*, *1*(3), 193-216.

Yang, Y. & Lui, X. (1999). A Reexamination of Text Categorization methods, In P*roceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, (pp. 42-49). University of California, Berkeley, USA.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases, In *Proceedings of the ACM SIGMOD Conference on Management of Data* (pp. 104-114). Montreal, Canada.