A New Method for Automated Categorizing and Finding Similarity of Online Persian News

Nasser Ezzati Jivan
n.ezzati@polymtl.ca
Mahlagha Fazeli
mfazeli@comp.iust.ac.ir
Khadije Sadat Yousefi
khyousefy@comp.iust.ac.ir

The Problem?

- Web pages need to be categorized
- There are many methods for English contents.
- Due to the different nature of Persian language, those methods aren't suitable for Persian.
- We are looking for a new method for categorizing Persian contents.

Web Page Categorization

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

Manual Categorization

- Web pages are analyzed and categorized with some experts in each field.
- Dmoz.org, Yahoo.com (Before 1998)

- High degree of precision
- Increasing number of web pages makes it very difficult and impractical

Cluster methods

- Extracting features
  - Each feature is a keyword or phrase appearing in a group of documents
- Each document is shown using a feature vector
- A clustering algorithm is used on the collection of vectors to categorize the documents.

Structural Categorization

- Structure of web pages is used widely to improve the organization search and analysis of information
  - A hyperlink shows the topical relationship between documents
- Anchor texts
- The text near a link
- Keywords
- …

Automatic Categorization of Persian Content

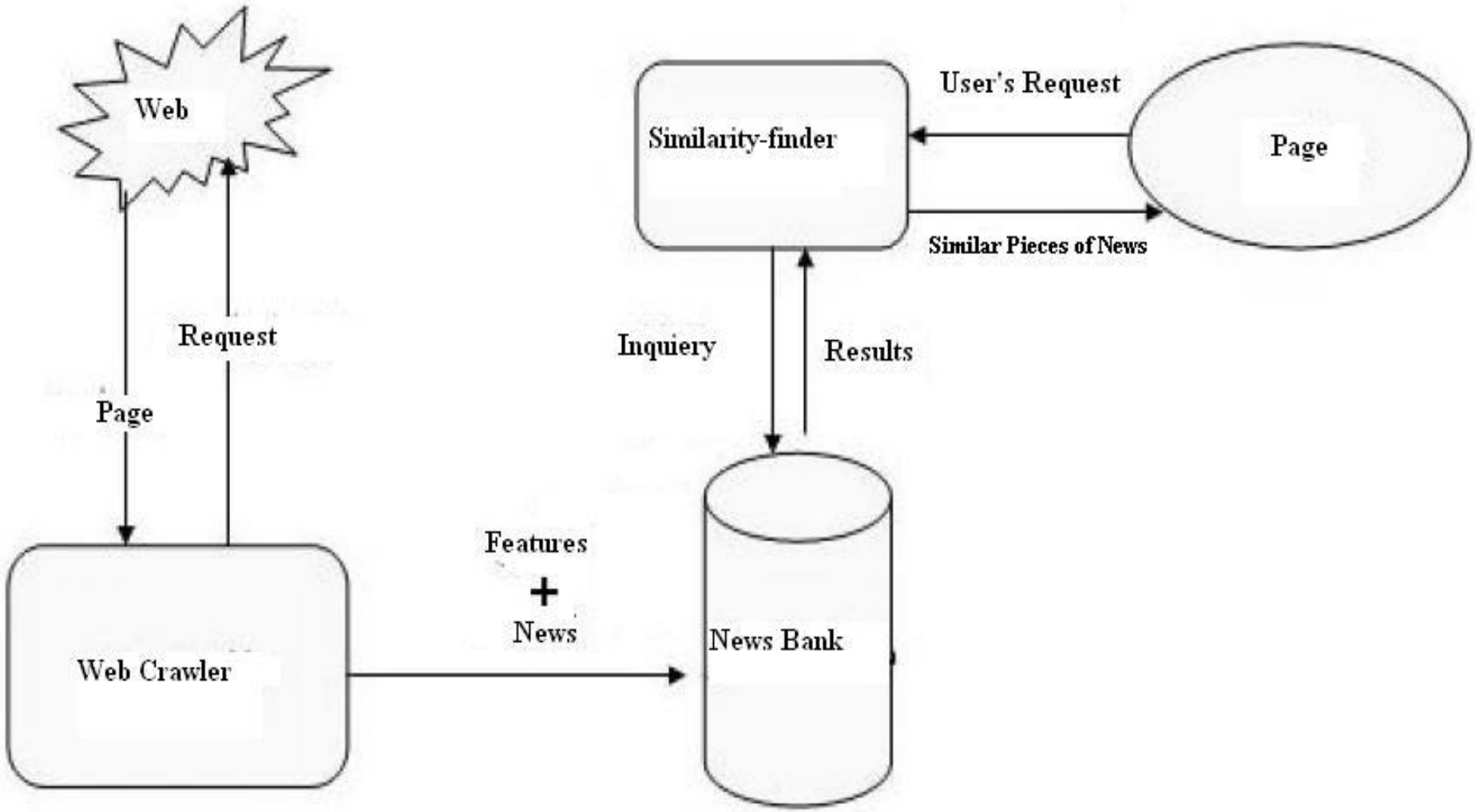
- We've Implemented and tested our method on Persian News.

The architecture of this approach includes five main modules:

- Pre-processing, Presentation, Storing, Categorizing and Similar finding and finally retrieval module
- Two main phases:

1- Feature Extraction

2- Categorizer and Similar Finder


Web
Similarity-finder
User's Request
Page
Similar Pieces of News
Request
Page
Inquiery
Results
Features
+
News
Web Crawler
News Bank

Extraction of features

- In this approach, we omit the stop(general) words.
  - The words don't have independent meaning.
  - Or they are very general and used in any sentences.
- Features are keywords of the news, the topics of the news, and their categories and news source and date.
- Therefore, those pieces of news which have similar features are regarded as related and similar

<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> → ":"|"?"|";"|","

Finding similarity of news

- The features of each piece of news are stored with the news in the news table.
- 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.

Finding similar pieces of news (2)

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 **n** th priority: all inquiries including only one of the keyword.

Result

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

Conclution

- Since in this system only the resemblance in subject and keywords in the news have been used, the results have about 80 % of accuracy.
- 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.
- Using of a thesaurus can improve the results.

Some Refrences

- M. Indra Devi, R. Rajaram, K. Selvakuberan, Generating best features for web page classification, Webology, Volume 5, Number 1, March.
- Chee Hong Chan ,Aixin Sun, Ee Peng lim , Automated Online News Classification with Personalization, Center for Advanced Information Systems, Nanyang Technological University.
- Chidanand Apte and Fred Damerau, Automated Learning of Decision rules for Text Categorization, ACM Transactions on Information Systems, Vol 12, No.3, pp.233-251.
- Dumais, S.T., Platt, J., Heckerman, D., and Sahami, M., Inductive Learning
- Algorithms and representations for text categorization, Proceedings of the Seventh
- International conference on Information and Knowledge Management pp.148-155.