Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 2214 – 2221

Procedia Engineering

www.elsevier.com/locate/procedia

International conference on modeling, optimization and computing

Customized News Filtering and Summarization System Based on Personal Interest

Anand babu M.H[a], Mani.G [b]

[a]Post-Graduate Scholar, Department of Computer science & Engineering, Anna University of Technology, Trichirapalli, India
[b]Assistant Professor, Department of Computer Science & Engineering, Anna University of Technology, Trichirapalli, India
[a] mhanandbabu@gmail.com, [b]gmani_it18@yahoo.co.in,

Abstract

Abstract— In the World Wide Web the information consists large amount of news contents. In Web intelligence Recommendation, filtering, and summarization of Web news have received much attention and also aims to find interesting news to the users .In this paper we present Customized news filtering and summarization system based on personal interest and it summarizes concise content for users .A user interest model induced by embedded learning component of CNFS and it also recommends customized news. Retrieving useful Web news involves both filtering and keyword extraction and also maintains key word knowledge base. The non-news content irrelevant to the news Webpage is filtered out .This extraction method substantially outperforms methods based on term frequency and lexical chains to represent semantic relation between words.

Keywords: ; Customized News; Web News Filtering; Web News Summarization;knowledgebase

1. Introduction

The web has started to play an important role for many people in delivering information related to their personal and professional lives. With the development of Internet, Web information has continued to proliferate at an exponential pace due to ease of publishing and access. News is the most popular interests for Web surfers. But the current Web pages always contain much irrelevant information such as advertisement links and navigation bar which is not interesting to the reader. So it is necessary to differentiate Web news content from others in Web pages. For retrieving useful web news it is important to accurately identify web news for correct filtering. With the rapid development of the World Wide Web, information on Web pages is rapidly inflated and congested with large amounts of news contents. The filtering and summarization of Customized Web news have drawn

doi:10.1016/j.proeng.2012.06.266

much attention in Web intelligence and also satisfies the user requirements. The filtering and summarization of Customized Web news refer to the recommendation, extraction, and summarization of interesting and useful information from Web pages. It is widely used to promote the automation degree in public opinion investigation, intelligence gathering and monitoring, topic tracking, and employment services

The first task of our system is to recommend interesting news to users. A news filter is applied in our system to provide high quality news content for analyzing. The second research component of the CNFS system is to summarize Web news. The summarization is given in the form of keywords based on lexical chains. Keywords offer a brief yet precise summary of the news content. Despite of the known advantages of keywords. A Web news recommendation mechanism is provided according to the users' interests which makes our CNFS system specially designed for Customized news treatment. An embedded learning component interacts with the recommendation mechanism and models users' interests. A keyword knowledge base is stored to update the users' profile. A keyword extraction algorithm is also provided to construct the lexical chains based on word sense disambiguation.

2. Related work

2.1. Web News Extraction

The targets of Web information extraction can be classified into three categories: records in a Web page, specific interesting attributes, and the main content of the page. Most Web information exploration systems for extracting records in a Web page work by automatically discovering record boundaries and then dividing them into items. The data to be extracted are often collocated in the same path of the DOM tree, and it is convenient to address data with DOM tree paths, which make the rule processing much easier. An extractive approach for title generation, which starts with URL tokens, HTML titles, keywords, and anchor text on incoming links etc. Their approach combines information from external sources, and performs probabilistic parameter learning with a URL's HTML title, context/abstract, and vocabulary at the source level. Web information extraction can be traced back to the integration research of heterogeneous data sources of structured and semi-structured data. A wrapper is viewed as a component in an information integration system to encapsulate accessing operations of multiple heterogeneous data sources, with which users can query on the integration system using a single uniform interface .
The NFAS system consists of two main tasks. Given a URL from an end user or an application, the
First task is to accurately identify whether the Web page is news or not, and if so filter the noise of the Web news, such as advertisements and non-relevant pictures. The second task is to summarize the Web news once it has been identified as a valid news page and has been filtered. The summarization is given in the form of lexical chains, based on keywords.

2.2. Keyword Extraction

Keyword extraction plays a key role in information retrieval, summarization, text clustering/classification, and so on. It aims at extracting keywords in terms of the text theme. In recent years, more and more web resources have been made available. Keyword extraction from web pages is helpful to deal with the subsequent web information extraction. At present, most of the news web pages have no keywords, and it is time-consuming and subjective to choose keywords.

Supervised extraction and unsupervised extraction. Supervised methods view keyword extraction as a classification task, where labelled keywords are used to learn a model. Naive Bayes to extract keywords, and designed the Kea system. Supervised methods are not very flexible because training on a

specific domain tends to customize the extraction process to that domain. Unsupervised keyword extraction removes the need for training data. Instead of trying to learn explicit features that characterize keywords, the unsupervised approach exploits the structure of the text itself to determine keywords that appear “central” to the text.

.

2.3**.** Lexical Chains

In text, lexical cohesion is the result of chains of related words that contribute to the continuity of lexical meaning. These lexical chains are a direct result of units of text being "about the same thing," and finding text structure involves finding units of text that are about the same thing. Hence, computing the chains is useful, since they will have a correspondence to the structure of the text. Determining the structure of text is an essential step in determining the deep meaning of the text. Lexical chains are used in many tasks, such as text retrieval and information extraction.

The construction of lexical chains needs a thesaurus for determining relations between words .Two thesauruses, including Word Net and How Net, are respectively used to compute word similarity in English and in Chinese. The word co-occurrence model is adopted to solve the problem that it is difficult to compute the semantic relations between words not in the thesaurus. The frequency of two words co-occurring in the same window unit (i.e., a sentence or a paragraph) can be computed without a thesaurus.

3. System architecture

A new user is required to register with an initial interesting topic category or keywords. Once a registered user logs in, the system returns Customized Web news to the user. When the user clicks on his/her interesting news items, the recently browsing history is updated. A keyword model is maintained to store the topic-distinguished keywords and the keywords selected from the browsed news stories.
The CNFS system consists of two phases.

Phase 1:

Customized Web News Filtering: In the customized news filtering phase one task is to filter out the news stories that are uninteresting to the users. Another is to filter out non-news parts on news Web pages. The Customized filtering subsystem has four components: a news aggregator, a news filter, a learning component, and a keyword knowledge base. The news aggregator automatically obtains content from news sources worldwide.

In this two learning algorithms including the k-nearest neighbor and Naïve Bayes are used to model the user’s preference and recommend Customized news. This filtering stage is accomplished by the Web Information Extractor that retrieves the news Web page’s title and news content by using pre-configured extraction rules. As with W4F ,the CNFS system also adopts extraction rules based on the paths of the DOM tree of the news Web page. The Web Information Extractor uses extraction Rules while it traverses the DOM tree of the Web page.

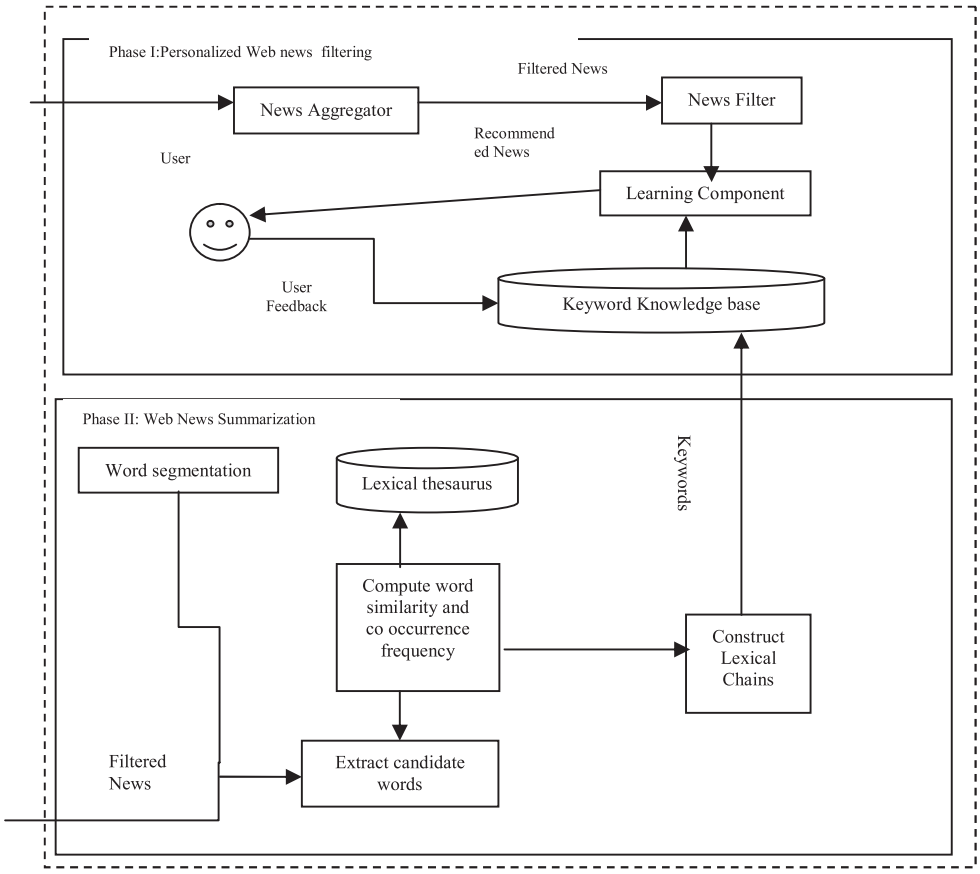

Fig1 Customized News Filtering and Summarization System

Customized news is being recommended by the learning component and simultaneously its learns the user interest model. The learning component interacts with the recommendation system, one way is to by the user recently browsed histories and the other way is to automatically select the keywords which allows user to modify.

.

Phase 2:

Web News Summarization:

During phase 1, we filtered non-Web news and non-news content on a news webpage. The next task is to summarize and extract the key phrases that capture the news web-page's main topic. We first segment the filtered document with a title and a body into words and then remove the stop words. Word Frequencies are counted and the TFIDF (term frequency–inverse document frequency) values are computed according to the corpus. Candidate phrases are identified by the TFIDF values. We also compute word co-occurrence frequencies and construct lexical chains with word similarities and word co-occurrence frequencies. Then key phrases are extracted from the candidate phrases according to the

TFIDF values and the semantic information in the lexical chains.

The filtered news content is segmented into words. Stop words are removed. Word frequencies are counted and the TFIDF values are computed according to the corpus. Candidate words are identified by the TFIDF values. For the candidate words that occur in the thesaurus, word similarities are computed. Word co-occurrence frequencies are also calculated. Lexical chains are constructed by word similarities and word co-occurrence frequencies. Then keywords are extracted from the candidate words according to the TFIDF values and the semantic information in the lexical chains.

4. Customized web news recommendation

Recommendation Algorithms: The *k*-nearest neighbor algorithm (*k*-NN) is a method for classifying objects based on closest training examples in the feature space. *K-NN* is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The *k*-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors (*k* is a positive integer, typically small).

Web news has several characteristics including dynamic content, changing interests, multiple interests, novelty, and so on, that make some approaches better suited than other approaches. In this paper, we focus on the content-based methods to recommend news by analyzing the user's browsing history. We divide the recommendation news into three groups: previous news tracking, interesting topics, and novelty news. We use the k-nearest neighbor algorithm to track previously read news and find novelty news. The k-nearest neighbor algorithm identifies recently known stories that the user has read. It keeps tracking new stories that have the same event thread with recently read stories, and it finds novel news. After filtering out the non-news parts on the Web news page, each news article is converted to a TFIDF. Although the k-nearest neighbour algorithm performs well in tracking news events and finding novel news, the recommended news stories are too specific that do not reflect the diversity of the user interests. Therefore, we use another probability learning model, Naïve Bayes to calculate the probability of news stories being interesting. Each news story is represented as a feature-value vector, where features are the keywords selected from the news story, and feature values are the word frequencies. For each news story (or the user preference), we can calculate the probability of the vector belonging to a given topic class according to the Naïve Bayes classifier.

Proposition 1 Assume that user u is independent to the news document d given the news topic classification model C = {c1,c2,...,cn} , where n is the number of news topic categories. The probability that document d is recommended to user u is computed as follows:

$$(u|d) = p(u) \sum_{j=1}^{n} \frac{p(c_j|u)p(c_j|d)}{p(c_j)}$$

Proof: According to the conditional probability formula, P(u | d) = p(u,d)/ p(d) , and by the total probability theorem.

Then $$p(u|d) = \sum_{j=1}^{n} \frac{p(u|c_j)p(d|c_j)p(c_j)}{p(d)}$$

Since $p(u|c_j)\,p(c_j) = p(u)\,p(c_j|u)$,and

$$p\{\{d|c_j\}|p(d)\} = p\{\{c_j|d\}|p(c_j)\},$$

$$P(u|d) = p(u)\sum_{j=1}^{n}\frac{p(c_j|u)p(c_j|d)}{p(c_j)}$$

For a given user ,p(u) is a constant value, so we can recommended d using the formula:

$$P(u|d) \propto p(u)\sum_{j=1}^{n}\frac{p(c_j|u)p(c_j|d)}{p(c_j)}$$

We formalize the recommendation algorithm as follows.

- FOR each upcoming news story DO
- calculate the similarities of the news story with the user's recently rated stories and get k most nearest neighbours.
- IF one of the k similarities is larger than t3 ;
- label the upcoming story as uninteresting;
- CONTINUE;
- IF the average of the k similarities is larger than t2 ;
- put the new story into the interesting queue;
- IF the average of the k similarities is less than t1 ;
- put the new story into the novelty queue;
- recommend the stories in the novelty queue in the ascending order of the average similarity;
- recommend the remaining stories according to the probability calculated by formula

Interaction of the Learning Component with the Recommendation System The evaluation of a recommendation system is a huge project that needs a long time to collect the users' data. The learning model is modified only when the performance of the recommendation system is evaluated. In the proposed CNFS system, the learning component is interactive with the overall system by the keyword knowledge and the user-click behaviours.

Keyword Extraction Algorithm

- Non-news content in the news Web page is filtered. Words are segmented and stemmed (for English words), and stop words are removed. Compute the TFIDF of each word using formula .
- Select the top n words by TFIDF as candidate words.
- Our system build the disambiguation graph in which each node is a candidate word that is divided into several senses (concepts), and each weighted edge connects two word senses.
- Perform the word sense disambiguation for each candidate word, and the one sense with the highest sum of similarities with other word senses is assigned to the word.
- Build the actual lexical chains. An edge connects two words if the word similarity exceeds the threshold t4 or the word co-occurrence frequency exceeds the threshold t5.
- Compute the weight of each candidate word.
- Select the top m words as the keywords extracted from the candidate words by their weights

Conclusion

This system provides users with efficient and reliable access to classified news from different sources. In this paper, we have presented the recommendation and summarization components of our customized news filtering and summarization (CNFS) system. For the recommendation component, we have designed a content-based news recommender that automatically obtains Word Wide Web news, and it also recommends personalized news to users according to their preference. Two learning strategies are used to model the user interest preference including the k-nearest neighbour and Naive Bayes. To better analyze the news content, a news filter is used to filter out the advertisements and other irrelevant parts on the news Web page

For the summarization component, a new keyword extraction method based on semantic relations has been presented in this paper. Semantic relations between words based on lexical thesaurus and word co-occurrence are studied, and lexical chains are used to link the relations. Keywords of high quality are extracted based on the information in the lexical chains.

References

[1]. A. Saiiuguet and F. Azavant, “Building intelligent web applications using light weight wrappers,” *Data and Knowledge Engineering*, 36(3): 283-316, 2001.

[2] G.Salton, A. Wong, and C. Yang, “On the specification of term values in automatic indexing,” *Journal of Documentation*, 29(4):351-372, 1973.

[3] I. H. Witten, G.W. Paynter, E. Frank, C. Gutwin, and C.G. Nevill- Manning, “KEA: Practical automatic key phrase extraction,” in *Proceedings of the 4th ACM Conference on Digital Libraries*, pages 254-256, Berkeley, California, US, 1999.

[4] X. Wu, G. Wu, F. Xie, Z. Zhu, X. Hu, H. Lu, and H. Li, “News filtering and summarization on the web,” *IEEE Intelligent Systems*, 25(5): 68-76, 2010.

[5] D.J. Hand and K. Yu, “Idiot's Bayes: not so stupid after all?” *Internat. Statist. Rev.* 2001, 69, 385-398.

[6] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl, “Group Lens: Applying collaborative filtering to Usenet news,” *Communications of the ACM 40*, 3: 77-87, 1997

[7] S. Li, H. Wang, S. Yu, C. Xin, “Research on maximum entropy model for keyword indexing,” *Chinese Journal of*

Computers, 27(9): 1192-1197, 2004

[8] Y. Liu, X. Wang, Z. Xu, B. Liu, "Ming constructing rules of Chinese key phrase based on rough set theory," *Acta Electronica Sinica*,35(2):371-374,2007