Information Retrieval Algorithm Based on Terminology Extraction and Context Based Searching Applied to WEB Pages

Prajakta Naik, JayeshPawar, AishwaryaBhide, Mandar Haldekar

nkprajakta@gmail.com, jayeshrpawar@gmail.com, aishwaryasbhide@gmail.com, mandar.haldekar@gmail.com

Pune Institute of Computer Technology, Pune-411043 ,India

Abstract. When browsing something on the Internet, a user is interested in quickly finding relevant information about the subject. This information can be extracted by identifying key ideas on the web-page. This paper discusses an approach to find relevant content from the web based on the keywords extracted from the document. Keywords are extracted by mining web page based on their context. Keywords are categorized in predefined domains such as place, books, people etc. Search is performed on web depending on context to improve the accuracy of retrieved information. Social and location-based information is integrated by leveraging social networks like Twitter and Facebook. Our proposal automates the task of finding related information about the web-page and thus reducing user's browsing efforts.

Keywords: Natural Language processing, Information Retrieval, Context-based searching, Machine Learning, Terminology extraction.

1 INTRODUCTION

With the huge and rapidly increasing amount of information available on the web, relevant information retrieval is one of the most challenging issue. When the user is reading a web document, very often he is interested in reading more about that subject and certain keywords related to that subject. The user normally submits a query to the search engine to find the information of his interest. The results may or may not be relevant to the current subject he is reading. This will consume extra search time to extract the relevant information he wants. Also if the topic of user's interest

contains images, videos and other information, the user needs to search different sites and read the contents. For example, consider the word “buffalo” that has come on different pages like “Wild Life Animals”, “Buffalo University”, “Buffalo Awards”. While reading any of these pages the user expects to get the relevant information about buffalo based on the context it is present in every page. Thus the user expects relevant information on the context of the word ‘buffalo’ used each time.

When a user is reading about a book, he is interested in knowing more about the book, its author, characters etc. Our proposal seeks to provide all such related information to the user on the same page without disturbing the current document the user is reading.

This approach provides the user with the information related to the current page through domain based extraction of keywords. Keywords are retrieved by natural language processing. These keywords are refined further according to the relevance of the subject using machine learning. Later they are classified under predefined domain. Context of each keyword is obtained by mining the web page. Finally the query based on the context is submitted to the search engine to get all relevant information like related news, images, video, tweets about that subject. Context based searching provides results according to the usage of the keyword in current document eliminating results of other possible contexts. This provides the user with useful information and reduces his effort of searching frequently on various sites.

The paper is organized as follows: Section 2 describes the terminology used. Section 3 gives the overview of our approach. Section 4 concludes the paper.

2 TERMINOLOGIES USED

Terminology extraction refers to the automatic extraction of keywords from a document. Natural Language Processing is used for this purpose. NP-chunks, POS tags and frequency distribution of words are obtained using NLP.

1. Frequency Distribution: With frequency distribution gives number of times particular word has occurred in the corpus.

2. NP-Chunks: The keywords of a document are mostly nouns or proper nouns or noun phrases with adjectives. These NP(noun phrase) chunks are source of useful information about the document.

3. Part-of-speech(POS) Tag : Part of Speech tagging is process of assigning part of speech to the words of corpus(text). It classifies corpus into noun, adjectives , verbs etc. and avails processing of text.

4. Collection frequency : It is the frequency of that word in the documents belonging to the same domain. Such collection of documents is stored for each predefined domain.

5. Position of the first of occurrence of the word: Generally, the keywords occur in the initial part of the document. So we can filter out the words that can most likely be keywords.

6. Context of keywords: In English, many words can be used in multiple distinct ways. A word can be a common noun, the name of a movie or a book, a proper noun etc. e.g. Emma can be a girl's name or it can refer to the famous novel by Jane Austen. The keywords are used in a specific sense in a document. This usage gives the context of the keyword as specific to the document.

3 System Overview

System for this proposal is given as,

$$S = \{Text, K, D, KD, SR, CT \mid \phi\} \quad (1)$$

Where, Text is the main content of current web page,

K is set of keywords , D is set of domains as:D={books,places,people,movies}, KD is set of domain related keywords. CT is collection of sets CT_i which are set of context for each $kd_i \epsilon$ KD .SR is set of search results. In this approach we will derive this system from web page which user is reading. User is interested in main text-content given in that document. Relevant

information retrieval is done in following step as shown in figure 1:

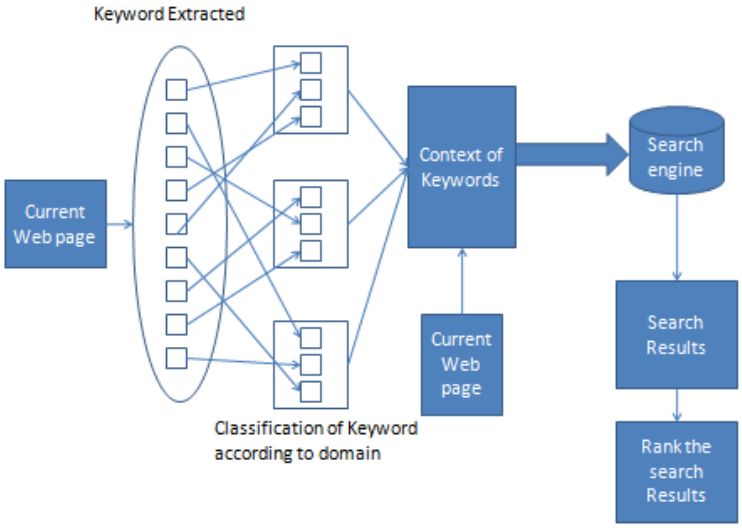

Fig. 1. Information Retrieval approach using terminology extraction and context based searching

1. Extraction of Keywords
2. Categorizing keywords into domains
3. Obtaining contexts of extracted keywords and performing web search using both keyword and context.
4. Extraction of the location based and social information from social networking sites.
5. Ranking Search Results

3.1 Extraction of Keywords

Keywords give overall idea of the document. So finding accurate keywords is very crucial. Web page scraping is done to process source code of page. Advertisement and comments on the page are discarded, while only main text content is retrieved and processed using NLP for keyword extraction.

First text content in web page is tokenized into words. All words are assigned POS tags. Since noun phrases contain maximum information, NP-chunks are obtained. Frequency distribution is applied to all NP-chunks. Collection frequencies and position of first occurrence of each NP-chunk are obtained. Based on all these factors, NP-chunks which can be keywords are found out. Machine learning is used to increase the efficiency of extracting keywords.

3.2 Categorizing Keywords into Pre-defined Domains:

Keywords extracted are filtered according to domains given by set D. Categorization of keywords into domains is done using domain relevance (DR) and domain consensus (DC) [1]

1.Domain Relevance:

It is measure of the degree of relevance of that term t to a specific domain D_i. DR is calculated as follows:

$$DR(t,D_i) = \frac{P(t|Di)}{\sum_{j=1..n} P(t|Dj)} \quad (2)$$

Where $P(t|D_i)$ is conditional probability of that term t being relevant to particular domain D_i . It is calculated as ratio of the frequency of that term t in domain D_i with total of its frequency in all domains of set D.

2.Domain Consensus:

It gives the measurement of distribution of term t in domain D_i. Distribution of t in a domain can be taken as random variable estimated throughout the corpus of that domain .

3.3 Obtaining Contexts of Extracted Keywords

Information Retrieval using only keywords returns large set of results many of which may not be relevant to the current web document. In our proposal, context of the keywords are found based on document title, tags and usage of those keywords in documents. We use context distance

method of standard vector space model [6]. This method measures closeness of word meanings. Context distance model measures semantic distance between two words using local context based on title, tags, other words in proximity etc .These contexts of keywords are considered while performing search on web. Augmented queries consisting of keywords and their context are formed and passed to search engine. It gives more relevant and comparatively precise results than normal keyword-based searching.

3.4 Extraction of Social and Location Based Information from Social Networking Sites

Search results are enriched by retrieving related information about user's friends on Facebook and followees from Twitter. e.g. When a person is reading about a book, he would like to have information whether any of their friends or followees have read or commented on that book. The information about the contents of the web page with respect to the user's location is also found out. e.g. If the user is reading a movie review, he might like to know about the movie theaters where that movie is currently playing.

3.5 Ranking the Search Results

With each search query, many results are returned by the search engine. These are re-ranked according to the relevance of search query. To rank the results ,tf-idf(term frequency -inverse document frequency) ranking method is used. Keyword plus its context are considered against corresponding search results. For each search query i.e.(keyword-context) , tf-idf weight is obtained. Results are sorted in decreasing order of tf-idf weights. In this way user can see most relevant results at the top of the search-result list.

4 Conclusion

This proposal presents an approach for retrieving relevant content about current web page using domain-based terminology extraction and context-based searching. Augmented search queries which include context of the automatically extracted keywords are used for searching. Therefore search

results are more satisfactory. This caters to the need of web users by automatically finding related useful information about web page.

References

1. Li Liu , Quan Qi, "Combined method for automatic domain-specific Terminology Extraction" Eight International conference on Fuzzy System and knowledge Discovery(FSDK)
2. Anette Hulth, "Improved Automatic Keyword Extraction given more linguistic knowledge "
3. Alexander Patry, Philippe Langlais ,"Corpus Based Terminology Extraction" 7th international conference of Terminology and knowledge Engineering, Copenhegan, Denmark, 2005, pp 313-321
4. Bhowmik, R. Sam, Huoston State Univesity, "Keyword Extraction from abstract and title" ,Southeaston, 2008 IEEE.
5. M. Brunzel and M spilopoulou, "Domain relevance on term weighting" Natural Language Processing and Information, pp 437-432, 2007
6. Hongyan Jing, Evelyne Tzoukermann, "Information Retrieval Based on Context Distance and Morphology", 22nd annual international ACM SIGIR conference on Research and development in information retrieval ,1999
7. Salton, G., & McGill, M. " Introduction to Modern Information Retrieval" McGraw Hill.