Joachim Griesbaum, Thomas Mandl, Christa Womser-Hacker (Hrsg.)

Information und Wissen: global, sozial und frei?

Proceedings des 12. Internationalen Symposiums für Informationswissenschaft (ISI 2011)

Hildesheim, 9.—11. März 2011

vwh
Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Constructing Topic-specific Search Keyphrase Suggestion Tools for Web Information Retrieval

Ari Pirkola

Department of Information Studies and Interactive Media
University of Tampere
Kanslerinrinne 1, Tampere 33014
ari.pirkola@uta.fi

Abstract

We devised a method to extract keyphrases from the Web pages to construct a keyphrase list for a specific topic. The keyphrases are identified and out-of-topic phrases removed based on their frequencies in the text corpora of various densities of text discussing the topic. The list is intended as a search aid for Web information retrieval, so that the user can browse the list, identify different aspects of the topic, and select from it keyphrases (e.g. find synonymous phrases) for a query. A keyphrase list containing a large set of keyphrases related to climate change was constructed using the proposed method. We argue that there is a need for such keyphrase suggestion tools, because the major Web search engines do not provide users with such terminological search aids that help them identify different topic aspects and find synonyms.

1 Introduction

The major Web search engines Bing, Google, Yahoo, and many others are necessary tools to find information from the Web, and they often provide users with good results. However, the users are often faced with the problem of finding such query keys that correctly represent their information needs. Formulating a good query requires that the user knows what aspects are re-

lated to the topic (s)he is interested in, so that (s)he can modify the query narrower or broader. As an example of an aspect, some of the aspects of *climate change* are glacier melting, sea-level rise, drought, adaptation, and political consequences – to mention a few among hundreds of aspects. Even though the user is interested in an aspect with which (s)he is familiar with, it is impossible to know all alternative expressions referring to the aspect used by the Web page authors. The user may use the query *sea level rise* but may lose the documents (s)he needs because in many relevant documents this concept is expressed differently, e.g. *rising sea level*, *rising seas*, or *higher sea level*. Moreover, authors often use elliptical expressions, i.e., phrases where one component is omitted (e.g. after introducing the full phrase *sea level rise* the author may refer to it by the elliptical phrase *the level rise*), and even such short forms may strengthen the query and affect document ranking positively.

Obviously, a list containing the most important phrases related to a particular topic would be an advantageous tool for Web searchers, helping to find good query keys. In this study, we devise a method to construct such a list, which is called *Topic-specific Search Keyphrase Suggestion Tool*. We are interested in scientific topics but the proposed method can be generalized to any reasonable topic. Here the *keyphrase* of the topic means a phrase that is often used in texts dealing with the topic and which refers to one of its aspects. The list is intended as an aid for Web information retrieval, so that the user can browse the list and select from it keys for a query. Each phrase in the list is assigned an *importance score* based on its frequencies in the text corpora of various *densities* of text discussing the topic. The keyphrases are extracted from pages relevant to the topic in question, and are thus known to appear in pages discussing the topic when used as search keys. Hence, the proposed approach implicitly involves the idea of reciprocity: keyphrases are extracted from relevant Web pages, and the phrases in turn can be used in queries to find relevant pages.

We encountered two main challenges when devising the keyphrase list: (1) How to identify pages that are relevant to the topic for use as keyphrase source data? (2) How to identify the keyphrases among all phrases in the relevant pages and prune out out-of-topic phrases?

In the first case, the method uses an information retrieval system to assign relevance scores to pages fetched by a focused crawler from the Web sites of universities and other research organizations investigating the topic. The

keyphrases of the topic are extracted from the pages assigned a high relevance score by the retrieval system.

Second, we introduce a novel method to identify keyphrases and to clean the keyphrase list from out-of-topic phrases. The method calculates importance scores for phrases on the basis of the frequencies of the phrases in the corpora of various densities of relevant text. The most obvious out-of-topic phrases receive a low importance score and are removed from the final list. An ideal case would be a large corpus that is dense in relevant text, but it is not easy to access large amounts of such texts. We therefore use a very dense corpus and an irrelevant corpus containing documents on a different topic than the topic for which the keyphrase list is constructed, and two corpora that are in-between these extremes. The dense corpora are built on the basis of the occurrences of the topic title phrase (e.g. climate change) and a few known keyphrases in the original corpus crawled from the Web. This approach allows us to separate between the keyphrases and out-of-topic phrases based on the fact that the relative frequencies of keyphrases decrease as the density decreases. After these automatic phases the list still contains some undesirable phrases which are removed manually (Section 3.2).

Using the proposed method, we constructed a keyphrase list for the topic *climate change*. The list is primarily intended for use in the scientific-based search system dealing with climate change (http://kastanja.uta.fi:8988/CLICS/) that was implemented in our earlier study, but it can be used as well together with general Web search engines to facilitate retrieving climate change related pages from the Web. The list is available on the Web at http://kastanja.uta.fi:8988/CLICS/about_index.html, and it contains 2533 two-word phrases and 848 three-word phrases.

The quality of the climate change keyphrase list was evaluated by determining (using samples) what proportion of the keyphrases and what proportion of all phrases in the crawled corpus (i.e., when keyphrase identification is not done) occur in the core content fields (title, abstract, keywords) of journal articles and conference papers dealing with climate change. The results showed that the proportion of keyphrases was higher than the proportion of phrases systematically selected from the crawled corpus.

2 Related Work

The proposed idea to construct a search keyphrase suggestion tool allowing searchers to see all important phrases related to a particular topic is novel. The new methodological idea behind our approach is to utilize the corpora of various densities of relevant text. Conventionally, keyphrase extraction refers to a process where phrases that describe the contents of a document are extracted and are assigned to the same document to facilitate e.g. information retrieval. Most conventional approaches are based on machine learning techniques. KEA (Witten et al., 1999), GenEx (Turney, 2003), and KP-Miner (El-Beltagy and Rafea, 2009) are three well-known keyphrase extraction systems. In these systems, keyphrases are identified and scored based on their length and their positions in documents, and using the TF-IDF weight.

Muresan and Harper (2004) also developed a terminological support for searchers' query construction in Web searching. However, unlike our study they did not focus on keyphrases but proposed an interaction model based on system-based mediation through structured specialized collections. The system assists the user in investigating the terminology and the structure of the topic of interest by allowing the user to explore a specialized source collection representing the problem domain. The user may indicate relevant documents and clusters on the basis of which the system automatically constructs a query representing the user's information need. The starting point of the approach is the ASK (Anomalous State of Knowledge) model where the user has a problem to solve but does not know what information is needed (Belkin et al., 1982). Lee (2008) showed that the mediated system proposed by Muresan and Harper (2004) was better than a direct IR system not including a source collection in terms of effectiveness, efficiency and usability. The more search tasks the users conducted, the better were the results of the mediated system.

We crawled the relevant documents from the Web sites of research organizations using a focused crawler. Focused crawlers are programs that fetch Web documents that are relevant to a pre-defined domain or topic (Hersovici et al., 1998; Diligenti et al., 2000; Pirkola and Talvensaari, 2010). Only documents assessed to be relevant by the system are downloaded and made accessible to the users e.g. through a digital library or a topic-specific search engine. During crawling link URLs are extracted from the pages and are added into a URL queue. The queue is ordered based on the probability

of URLs (i.e., pages pointed to by the URLs) being relevant to the topic in question. Pages are assigned probability scores e.g. using a topic-specific terminology, and high-score pages are downloaded first. Focused crawling research has focused on improving crawling techniques and crawling effectiveness (Diligenti et al., 2000; Bergmark et al., 2002; Pirkola and Talvensaari, 2010), and we are not aware of any study investigating the use of focused crawling for keyphrase extraction. Perhaps the closest work to our research is that of Talvensaari et al. (2008) who also constructed word lists using focused crawling. However, they used focused crawling as a means to acquire German-English and Spanish-English comparable corpora in biology for statistical translation in cross-language information retrieval.

3 Methods

3.1 The Crawler

We implemented a focused crawler in which the relevance of the pages during crawling is determined by matching a topic-defining query against the retrieved pages using a search engine. We used the Lemur search engine (http://www.lemurproject.org/) which allows the use of a proximity operator and weighted queries. The topic-defining query contained the following query keys: #3(climate change), #3(climate research), climate, climatic, #3(research project), research. The words combined by the proximity operator #3 are not allowed to be more than three words apart from each other to match. The keys were combined by Lemur's weighted #sum operator to give more weight to the first keys above than the last two keys that relate to research activity in general. The pages with relevance scores higher than a given threshold were kept in the crawling results. The irrelevant corpus was crawled similarly to the relevant corpus, except that now we fetched Web documents on *genetics*, and the topic-defining query contained genetics related words and phrases. In all, crawling gave some 3100 documents deemed to be relevant and some 3600 irrelevant documents.

A focused crawler does not follow all links on a page but it will assess which links to follow to find relevant pages. Our crawler assigns the prob-

ability of relevance to an unseen page v using the following formula, which gave good results in a preliminary experiment.

$\Pr(T|v) = (\alpha * \text{rel}(u) * (1/\log(N_u)) + ((1 - \alpha) * \text{rel}(<u,v>))$,

where α is a weighting parameter ($0 < \alpha < 1$), rel(u) is the relevance of the seen page u, calculated by Lemur, N_u the number of links on page u, and rel(<u,v>) the relevance of the link between *u* and the unseen page *v*. The relevance of the link is calculated by matching the *context* of the link against the topic query. The context is the anchor text, and the text immediately surrounding the anchor. The context is defined with the help of the Document Object Model (DOM): all text that is within five DOM tree nodes of the link node is considered belonging to the context. The Document Object Model is a convention for representing and interacting with objects in HTML, XHTML and XML documents (http://en.wikipedia.org/wiki/Document_Object_Model).

As can be seen, $\Pr(T|v)$ is a sum that consists of two terms: one that depends on the relevance of the page, and one that depends on the relevance of the link. The relative importance of the two terms is determined by the weight α. Based on our crawling experiment we selected for the α parameter we used the value of $\alpha = 0.3$. Also, the number of links on page *u* inversely influences the probability. If rel(*u*) is high, we can think that the page "recommends" page *v*. However, if the page also recommends lots of other pages (i.e., N_u is high), we can rely less on the recommendation.

3.2 Constructing the Climate Change Keyphrase List

We now describe how the climate change keyphrase list was constructed. In the first phase, Web pages dealing with climate change were crawled using the focused crawler described in Section 3.1. The start URL set contained some 80 URLs of the most productive organizations engaged in climate change research, which were identified using the Scopus citation database (http://www.scopus.com/). The crawling scope of the crawler was limited so that the crawler was only allowed to visit the pages on these start sites, and their subdomains (for example, *research.university.edu* is a subdomain of *www.university.edu*), as well as sites that are one link apart from the start domain. These restrictions ensured that the crawling results do not degrade but crawling keeps in scientific sites.

The first phase of the *processing* of the crawled data was to extract all bigrams (i.e., two consecutive words) and trigrams (i.e., three consecutive words) from the crawled relevant and irrelevant corpora, and to recognize which bi- and trigrams are phrases. For phrase identification we used the small word (stop-word) technique (Jaene and Seelbach, 1975) and kept those bi- and trigrams only that were surrounded by small words and that did not include a small word. The small word list was a standard stop-word list of an information retrieval system, and it contained 856 words. In scientific documents words related to research (e.g. study, author) are intermixed with the keyphrases of the topic, and the best way to remove them is to handle them as if they were stop-words. We therefore supplemented the list with a small set of research-related words (N=18), and removed all phrases that include such a word.

In this study, we introduce a novel method to identify keyphrases and to clean the keyphrase list from out-of-topic phrases. The crawled relevant corpus was divided into three separate corpora based on the occurrences of the topic title phrase (climate change) and a few known keyphrases related to climate change. The three corpora differ from each other in the density of text portions containing keyphrases. We first identified the known keyphrases (N=10) that well represent the topic, such as *global warming* and *sea level*. The three corpora were as follows: (1) The whole relevant corpus; (2) A corpus where each text line contains the topic title phrase; (3) A corpus where each text line contains, in addition to the topic title phrase, at least one of the known keyphrases. The second corpus is denser in relevant text portions than the whole corpus, and the third one is denser than the second. The fourth corpus was the irrelevant corpus, which obviously only contains a few keyphrases. It can also be assumed that the frequency of keyphrases is relatively higher in the third corpus than in the first two corpora. The second corpus, in turn, is assumed to contain relatively more keyphrases than the first one. Out-of-topic phrases occur in the irrelevant corpus. They can be expected to be infrequent in the dense corpora simply because there is not much room for them in text portions that have many keyphrases.

After these automatic phases, the list still contained some undesirable phrases, in particular non-specific phrases (such as *take action*) and phrases containing non- or weakly informative verbs (such as *addressing climate change*). These phrases were removed manually. Generally, the number of removals depends on the applied importance score (Section 3.3) threshold. All phrases in the crawled corpus could be assigned an importance score and

if a very low threshold would be applied, the percentage of removals would be high. In the case of a high threshold none or only a few phrases need to be removed.

3.3 Importance Score

Below we introduce notational conventions used in the importance score calculations.

Notational Conventions.

Let P_2 be some two-word phrase in the first document corpus *DC(1)*, i.e., $P_2 \in DC(1)$. We denote its frequency in the corpus by $F_{DC(1)}(P_2)$. Correspondingly, the frequency of a three-word phrase in the first corpus is denoted by $F_{DC(1)}(P_3)$. The frequencies of the two- and three-word phrases in the second, third and fourth corpora are denoted similarly, e.g. three-word phrases in the fourth corpus: $F_{DC(4)}(P_3)$.

Assumedly, a phrase which has a high frequency in the three relevant corpora and a low frequency in the fourth corpus deserves a high score. Therefore, the importance score for the two- and three-word phrases is calculated as follows (in the calculations the value 0 is converted into 1):

$IS(P_2) = \ln(F_{DC(1)}(P_2) * F_{DC(2)}(P_2) * F_{DC(3)}(P_2) / F_{DC(4)}(P_2))$

$IS(P_3) = \ln(F_{DC(1)}(P_3) * F_{DC(2)}(P_3) * F_{DC(3)}(P_3) / F_{DC(4)}(P_3))$

Table 1 shows the 20 highest ranked two- and three word phrases in the climate change keyphrase list and their importance scores. The whole list is available at http://kastanja.uta.fi:8988/CLICS/about_index.html. Most of the keyphrases are established phrases. The phrase *change impacts* is an example of an elliptical phrase. Such short forms are understandable in the context of climate change, and as argued in Section 1 they may be good query keys.

Table 1. The highest ranked keyphrases in the climate change keyphrase list.

Two-word phrases	$IS(P_2)$	Three-word phrases	$IS(P_3)$
change impacts	19.0	climate change impacts	18.5
greenhouse gases	17.9	greenhouse gas emissions	17.5
global warming	17.6	climate change adaptation	16.6
climate changes	16.8	future climate change	16.3
future climate	16.5	global climate change	16.2
greenhouse gas	15.9	climate change projections	16.2

Two-word phrases	IS(P_2)	Three-word phrases	IS(P_3)
carbon dioxide	15.7	global environmental change	16.1
global climate	15.5	fourth assessment report	15.7
adaptation strategies	15.4	climate change issues	14.4
earth system	15.3	climate change mitigation	14.3
potential impacts	15.3	induced climate change	14.2
greenhouse effect	15.2	sea level rise	14.1
food security	15.0	climate change scenarios	13.6
climate adaptation	15.0	regional climate change	13.5
sustainable development	15.0	climate change policy	13.4
climate policy	14.8	dangerous climate change	12.7
potential impact	14.7	abrupt climate change	12.6
climate action	14.6	climate change report	12.3
climate system	14.6	climate change program	12.2
ozone layer	14.4	greenhouse gas concentrations	12.2

4 Evaluation

The quality of the climate change keyphrase list was evaluated by determining (1) what proportion of the keyphrases in the list (test situation) and (2) what proportion of phrases selected from the relevant corpus (baseline situation) occur in the core content fields (title, abstract, keywords) of journal articles and conference papers dealing with climate change. In the first case, a systematic sample of keyphrases (N=50 both for two- and three-word phrases) was selected from the keyphrase list. In the second case, similarly to the first case, a systematic sample of two- and three-word phrases (N=50 for both) was selected from the relevant corpus (containing both keyphrases and out-of-topic phrases). If the proposed method effectively identifies keyphrases, as is expected, their proportion will be considerably higher than that of corpus phrases. On the other hand, of all the corpus phrases a large proportion is keyphrases, so they are not infrequent in the core fields of relevant articles and papers. Hence, the main question in the evaluation is whether the second stage of the proposed approach (the use of corpora of various densi-

ties of relevant text) improves the effectiveness compared to the first stage alone (constructing the relevant corpus by means of focused crawling).

In this evaluation experiment we used the Web of Science citation database (http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science). In the Web of Science, each journal article and a conference paper is represented by a record that contains the core content fields *title*, *abstract*, and *keywords* and several other fields. The query used in the evaluation was expressed as follows: Find documents where the keyphrase (test situation) / corpus phrase (baseline situation) and the topic title phrase (climate change) occur in the same record in the title, abstract, or keyword field. For example, we searched for documents that contain in their core fields both the phrase *abrupt change* and the title phrase *climate change*.

The results of the evaluation experiment are reported in Table 2. As described above, in each four cases we selected 50 phrases, and column 2 shows how many of them occur in the core fields of articles and papers together with the phrase *climate change*. As shown, the number of keyphrases is remarkably higher than that of corpus phrases. In the case of two-word keyphrases, all 50 have occurrences (at least one occurrence) whereas only 26 two-word corpus phrases have occurrences. Column 3 indicates the total number of occurrences for the 50 keyphrases and for the 50 corpus phrases. Column 4 indicates the average number of occurrences per keyphrase and per corpus phrase. For keyphrases the total number of occurrences and the average are considerably higher. Two-word keyphrases appear more frequently than three-word keyphrases.

Table 2. *The results of the evaluation experiment.*

Phrase type	N:o keyphrases; N:o corpus phrases	N:o occurrences	Average n:o occurrences
2-word keyphrases	50	11 992	239,8
2-word corpus phrases	26	609	12,2
3-word keyphrases	43	2 743	54,9
3-word corpus phrases	14	184	3,7

5 Conclusions

Conventionally, keyphrase extraction refers to a process where phrases that describe the contents of a document are extracted and are assigned to the same document to facilitate e.g. information retrieval. We presented a novel approach which differs from the conventional approach in that we do not handle individual documents but a set of documents discussing a particular topic. From these documents we extract keyphrases that describe different aspects of the topic. The proposed method is based on the use of several document corpora of different densities of relevant text.

Our project plan involves building a multi-topic search keyphrase suggestion tool dealing with many globally significant topics. The climate change keyphrase list will be a part of the larger tool. We believe that such a multi-topic tool is needed in scientifically-oriented Web information retrieval. It will serve users such as researchers and journalists searching for information on scientifically and globally important information. It may also be possible to apply the keyphrase list in areas other than information retrieval (e.g. document clustering), which may be one direction of our future research.

Acknowledgments

This study was funded by the Academy of Finland (research projects 130760, 218289).

References

Belkin, N. J., Oddy, R. N., Brooks, H. M. (1982). ASK for information retrieval: Part I. Background and history. Journal of Documentation, 38 (2), pp. 61–71.

Bergmark, D., Lagoze, C., Sbityakov, A. (2002). Focused crawls, tunneling, and digital libraries. Sixth European Conference on Research and Advanced Technology for Digital Libraries, Rome, Italy, September 16–8, pp. 91–106.

Diligenti, M., Coetzee, F. M., Lawrence, S., Giles, C. L., Gori, M. (2000). Focused crawling using context graphs. Twenty-sixth International Conference on Very Large Databases (VLDB), pp. 527–534.

El-Beltagy, S. and Rafea, A. (2009). KP-Miner: A keyphrase extraction system for English and Arabic documents. Information Systems, 34(1), pp. 132–144.

Hersovici, M., Jacovi, M., Maarek, Y., Pelleg, D., Shtalhaim, M., Ur, S. (1998). The shark-search algorithm – an application: tailored Web site mapping. Seventh International Conference on World Wide Web, Brisbane, Australia.

Jaene, H. and Seelbach, D. (1975). Maschinelle Extraktion von zusammengesetzten Ausdrücken aus englischen Fachtexten. Report ZMD-A-29. Beuth Verlag, Berlin.

Lee, H. J. (2008). Mediated information retrieval in Web searching. Proceedings of the American Society for Information Science and Technology, 45(1), pp. 1–10.

Muresan, G. and Harper, D. J. (2004). Topic modeling for mediated access to very large document collections. Journal of the American Society for Information Science and Technology, 55 (10), pp. 892–910.

Pirkola, A. and Talvensaari, T. (2010). Addressing the limited scope problem of focused crawling using a result merging approach. Proceedings of the 25th Annual ACM Symposium on Applied Computing (ACM SAC), Sierre, Switzerland, March 22–6, pp. 1735–1740.

Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., Laurikkala, J. (2008). Focused Web crawling in the acquisition of comparable corpora. Information Retrieval, 11(5), pp. 427–445.

Turney, P. D. (2003). Coherent keyphrase extraction via Web mining. Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico, pp. 434–439.

Witten, I. H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C. G. (1999). KEA: Practical automatic keyphrase extraction. Proceedings of the 4th ACM conference on Digital Libraries, Berkeley, California, pp. 254–255.