TOPIC-SPECIFIC WEB SEARCHING BASED ON A REAL-TEXT DICTIONARY

Ari Pirkola
University of Tampere, School of Information Sciences, Tampere, Finland

Keywords: Dictionaries, Focused Crawling, Query Performance Prediction, Searching, Vertical Search Engines, Web Search Engines.

Abstract: The contributions of this paper are twofold. First, we present a new type of dictionary that is intended as a search assistance in topic-specific Web searching. The method to construct the dictionary is a general method that can be applied to any reasonable topic. The first implementation deals with climate change. The dictionary has the following new features compared to standard dictionaries and thesauri: (A) It contains real-text phrases (e.g. rising sea levels) in addition to the standard dictionary forms (sea-level rise). The phrases were extracted automatically from the pages dealing with climate change, and are thus known to appear in the pages discussing climate change issues when used as search terms. (B) Synonyms, i.e., different spelling, syntactic, and short form variants of the phrase are grouped together into the same entry (synonym set) using approximate string matching. (C) Each phrase is assigned an importance score (IS) which is calculated based on the frequencies of the phrase in relevant pages (i.e., pages on climate change) and non-relevant pages. Second, we investigate how effective the IS is for indicating the best phrase among synonymous phrases and for indicating effective phrases in general from the viewpoint of search results. The experimental results showed that the best phrases have higher ISs than the other phrases of a synonym set, and that the higher the IS is the better the search results are. This paper also describes the crawler used to fetch the source data for the climate change dictionary and discusses the benefits of using the dictionary in Web searching.

1 INTRODUCTION

In a current research project, we developed a vertical search engine (a topic-specific search engine) (Pirkola, 2011b) and a topic-specific dictionary of *real-text phrases*, that is, a dictionary of phrases that are extracted from the Web pages discussing the topic in question. Both are dedicated to *climate change* and their focus is on scientific pages, but the methods to implement the search engine and the dictionary are general methods that can be applied to other topics. The search engine and the dictionary are at: www.searchclimatechange.com.

The proposed real-text dictionary, in short RT-dictionary, has new features compared to the standard dictionaries and thesauri. We now discuss the existing climate change RT-dictionary, but a similar dictionary can be constructed for any reasonable topic using the method presented in this paper. The dictionary contains some 5 500 unique real-text phrases related to climate change (*keyphrases*). They were extracted from scientific Web pages discussing climate change. Each phrase is assigned a frequency-based *importance score*, which reflects the significance of the phrase in the context of climate change research. Different variant forms of the same phrase, such as *sea-level rise*, *sea level rising*, and *rising sea level*, are grouped together into the same entry (synonym set) using approximate string matching. The dictionary was developed for use as a search assistance in our search system to support query formulation, but it can be used as well together with the general Web search engines. The phrases represent different aspects of the climate change research. When used in searching the user can browse through the dictionary to clarify his or her information need, or select more directly appropriate search phrases.

The purpose of the importance score (IS) is to provide the user with information on important and less important phrases. Within a synonym set all phrases are not equal from the viewpoint of the

Pirkola A..
TOPIC-SPECIFIC WEB SEARCHING BASED ON A REAL-TEXT DICTIONARY.
DOI: 10.5220/0003895602890298
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 289-298
ISBN: 978-989-8565-08-2

search results, but some yield better results than others. It may be that one is significantly better than the other phrases. In the experimental part of this paper we investigate, first, how effective the IS to indicate the best phrase within a synonym set. We expect the phrase with the highest IS to yield the best search results. Our second research question is: Does high IS mean high retrieval performance in general (i.e., when synonymy is not considered). The expectation is that the higher the IS, the better the search results. The results of this study guide our further work: depending on the results we may elaborate the calculation of the IS. The results may also suggest a sensible lower limit of the IS to be applied in the dictionary. The results also guide more generally our work in the further development of the RT-dictionary.

In summary, this paper has two contributions. First, we describe how the RT-dictionary of climate change was constructed (Section 4). The presented methods and tools can be used to a construct a similar RT-dictionary for any reasonable topic. Second, in the experimental part of this paper we investigate how effective the proposed importance score is to indicate the best phrase among synonymous phrases and to indicate effective phrases in general from the viewpoint of the search results (Sections 5 and 6). We also review related research (Section 2) and describe the crawler used to fetch the source data for the climate change RT-dictionary (Section 3). Section 7 presents the discussion, and Section 8 concludes this paper.

2 RELATED WORK

This study is related to research on query performance prediction, which is an important research problem in the field of information retrieval research. A retrieval system that can predict the difficulty of a query can adapt its parameters to the query, or it can give feedback to the user to reformulate the query. Query performance prediction has been the focus of many studies. He and Ounis (2006) studied six methods to predict query performance, e.g. average inverse collection term frequency and the standard deviation of inverse document frequency. Cronen-Townsend et al. (2002) proposed computing divergence of the query model from the collection model to predict the difficulty of queries. Perez-Iglesias and Araujo (2010) proposed the use of the maximum standard deviation in the ranked-list as a retrieval predictor. Different from these studies the IS ranks phrases with respect to each other rather than tries to predict retrieval performance as such.

This work is also related to research on keyphrase extraction. 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. Our approach differs from the conventional approach in that we do not handle individual documents but a set of documents discussing a particular topic. 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 keyphrase extraction systems presented in the literature. 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 retrieval 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.

The source data for the climate change RT-dictionary were crawled from the Web sites of universities and other research organizations investigating climate change by the focused crawler described in Section 3. Focused crawlers are programs that fetch Web documents (pages) that are relevant to a pre-defined domain or topic. 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 vertical 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 (Bergmark et al., 2002; Chakrabarti et al., 1999; Diligenti et al., 2000; Tang et al., 2005). Below we consider these papers. 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.

Bergmark et al. (2002) use the term tunneling to refer to the fact that sometimes relevant pages can be fetched only by traversing through irrelevant pages. A focused crawling process using the tunneling technique does not end immediately after an irrelevant page is encountered, but the process continues until a relevant page is encountered or the number of visited irrelevant pages reaches a pre-set limit. It was shown that a focused crawler using tunneling is able to fetch more relevant pages than a focused crawler that only counts relevant pages. However, efficiency is lowered due to the downloaded irrelevant pages.

Chakrabarti et al. (1999) utilized the Document Object Model (DOM) of Web pages in focused crawling. The DOM is a convention for representing and interacting with objects in HTML, XHTML and XML documents (http://en.wikipedia.org/wiki/ Document_Object_Model). A DOM tree represents the hierarchical structure of a page: the root of the tree is usually the HTML element which typically has two children, the HEAD and BODY elements, which further are divided into sub-elements. The leaf nodes of the DOM tree are typically text paragraphs, link anchor texts, or images. In addition to the usual bag-of-words representation of Web pages, the approach proposed by Chakrabarti et al. (1999) represented a hyper-link as a set of features <t, d> where t is a word appearing near the link, and d its distance from the link. The distance is measured as the number of DOM tree nodes that separates t from the link node. These features were used to train a classifier to recognize links leading to relevant pages. The links with low distance to relevant text are considered to be more important than links that are far from the relevant text.

In the Context Graph approach (Diligenti et al., 2000) a graph of several layers depth is constructed for each page and the distance of the page to the target pages is computed. In the beginning of crawling a set of seed URLs is entered in the focused crawler. Pages that point to the seed URLs, i.e., parent pages, and their parent pages (etc.) form a context graph. The context graphs are used to train a classifier with features of the paths that lead to relevant pages.

Tang et al. (2005) built a focused crawler for mental health and depression. The aim was to identify high-quality information on these topics. They found that link anchor text was a good indicator of the page's relevance but not of quality. They were able to predict the quality of links using a relevance feedback technique.

3 CRAWLER

We developed a focused crawler that is used to fetch pages both for the climate change search system and for use as source data for the RT-dictionary of climate change. The crawler can be easily tuned to fetch pages on other topics. In this section we describe the crawler.

The crawler determines the relevance of the pages during crawling by matching a topic-defining query against the retrieved pages using a search engine. It uses the Lemur search engine (http://www.lemurproject.org/) for this purpose. The pages on climate change were fetched using the following search terms in the topic-defining query: *climate change*, *global warming*, *climatic change*, *research*. We used the core journals in the field to find relevant start URLs. When pages on some other topic are fetched only the search terms and the start URL set need to be changed. So, applying the crawler to a new topic is easy.

To ensure that the crawler fetches mainly scholarly documents its crawling scope is limited, so that it is only allowed to visit the pages on the start URL 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. If needed, this restriction can be relaxed so that the crawling scope is not limited to these sites.

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 probability of relevance to an unseen page v using the following formula that was developed in our previous study. We call it *N-formula*:

$$Pr(T|v)= (\alpha * rel(u) * (1/\log(Nu))) + ((1 - \alpha) * rel(<u,v>)), \ \alpha= 0.3$$

where Pr(T|v) is the probability of relevance of the unseen page v to the topic T, α is a weighting parameter ($0 < \alpha < 1$), rel(u) is the relevance of the seen page u, calculated by Lemur, Nu 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: all text that is within five DOM tree nodes of the link node is considered belonging to the context.

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 parameter α. 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., Nu is high), we can rely less on the recommendation.

The N-formula gave the best results in an experiment where we compared it to two reasonable baseline methods. In the first baseline, link <u,v> was assigned the same relevance score as page u. In the second one, a link was assigned a relevance score based on its context. Crawling results were measured as the number of obtained relevant pages when the same number of documents was downloaded.

We also conducted an experiment where the weighting parameter α was set to (A) α=0.3 and (B) α=0.7. In (A) the crawler fetched about two times more relevant documents than in (B). Based on this crawling experiment we apply for the α parameter the value of α = 0.3.

4 REAL-TEXT DICTIONARY OF CLIMATE CHANGE

Constructing the real-text dictionary of climate change involved two main stages: (1) extraction of keyphrases and the calculation of importance scores and (2) identification of synonyms. Section 4.1. describes the first stage and Section 4.2 the second one. Section 4.2 also describes the synonym types contained in the dictionary.

4.1 Keyphrase Extraction

Here *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. Keyphrases typically represent different research areas of the topic (i.e., climate change in this study), some are more technical in nature. The keyphrase extraction method and a climate change *keyphrase list*, a predecessor of the climate change dictionary. are described in (Pirkola, 2011a**)**. Here we describe the main features of the extraction method, and present the importance score.

The source data for the dictionary, i.e., pages related to climate change, were crawled from the Web sites of universities and other research organizations investigating climate change by the crawler described in Section 3. When the crawled data were processed the first challenge was to determine which sequences of words are *phrases*. Here we applied the phrase identification method by Jaene and Seelbach (1975). The main point of the technique is that a sequence of two or more consecutive words constitutes a phrase if it is surrounded by small words (such as *the*, *on*, *if*) but do not include a small word (except for *of*). The phrases were extracted from the pages assigned a high relevance score by the crawler.

When constructing RT-dictionary the importance score is needed to prune out out-of-topic phrases from the dictionary. The most obvious out-of-topic phrases receive a low score and are not accepted in the dictionary. The remaining phrases are regarded as keyphrases and are included in the dictionary. This is the first application of the IS. The second one, which is investigated in this paper, is to use the IS as an indicator of the effectiveness of different keyphrases in searching.

The IS is calculated on the basis of the frequencies of the phrases in the corpora of various densities of relevant text, and in a non-relevant corpus. We determined the IS using four different corpora. The relevant corpora are built on the basis of the occurrences of the topic title phrase (i.e., climate change) and a few known keyphrases related to climate change in the original corpus crawled from the Web. Assumedly, a phrase which has a high frequency in the relevant corpora and a low frequency in the non-relevant corpus deserves a high score. Therefore, the importance score is calculated as follows:

$$IS(P_i) = \ln(F_{DC(1)}(P_i) * F_{DC(2)}(P_i) * F_{DC(3)}(P_i) / F_{DC(4)}(P_i)); \quad (F_{DC} > 0)$$

$F_{DC(1)}(P_i)... F_{DC(4)}(P_i)$ = the frequencies of the phrase P_i in the four corpora.

DC(1) = Highly dense corpus
DC(2) = Very dense corpus
DC(3) = Dense corpus
DC(4) = non-relevant corpus

This method allows us to indicate the importance of the phrase in the Web texts discussing climate change (or any topic for which RT-dictionary is constructed), and separate between keyphrases and out-of-topic phrases based on the fact that the relative frequencies of the keyphrases decrease as the density decreases.

In the dictionary, the importance scores range from 4.7-35.8. The choice of the lower limit of 4.7 has no experimental basis. As mentioned, one purpose of this study is to consider this issue.

Table 1 shows the 20 highest ranked phrases in the dictionary and their importance scores. Many of the phrases are well-known in the context of climate change research. The dictionary also contains short form phrases, i.e., phrases where one component is omitted. Authors often use such forms. For example, after introducing the full phrase *climate change impacts* the author may refer to it by the phrase *change impacts*. In searching such short forms may strengthen the query and affect document ranking positively when used together with their full forms. However, the short forms are often ambiguous, and often it does not make sense to use a short form without at the same time using its full phrase.

4.2 Synonyms

Synonyms were identified using the digram approximate matching technique. Phrases were first decomposed into digrams, i.e., substrings of two adjacent characters in the phrase (for n-gram matching, see Robertson and Willett, 1998). The digrams of the phrase were matched against the digrams of the other phrases in the phrase list generated in the keyphrase extraction phase (Section 4.1). Similarity between phrases was computed using the Dice formula, and the phrase pairs that had the similarity value higher than the threshold (SIM=0.75) were regarded as synonyms. The output of the digram matching process contained some 10% of wrong matches which were removed manually. In the majority of cases the identification of the wrong matches was a trivial task and it does not require any specific expertise. Currently, we are working to develop a more effective synonym identification method where only a minimal manual intervention is needed. We are testing several approximate string matching methods in combination with stemming and as such, as well as the use of the textual contexts of the phrases.

Table 1: Highest ranked keyphrases in the climate change RT-dictionary.

Keyphrases	IS
climate change	35.8
climate change research	28.4
global warming	26.7
impacts of climate change	26.3
sea level	26.2
global climate	25.7
greenhouse gas	25.3
sea level rise	25.2
climate change impacts	25.0
sustainable communities	25.0
environmental policy	24.9
environmental change	24.7
global climate change	24.5
sustainable energy	24.5
climate impacts	24.3
integrated assessment	24.2
water resources	24.1
gas emissions	23.9
greenhouse gases	23.9
climate changes	23.8

The dictionary contains the following types of synonyms:

A. Spelling Variants. Phrases that contain the same component pairs but are written slightly differently, e.g. *climate change prediction - climate change predictions*. These kinds of variants are mostly morphological variants.

B. Syntactic Variants. Phrases that contain the same component pairs, but the order of the components is different, e.g. *predicted climate change - climate change predictions*. These kinds of variants may also include type A variation, as in the example above.

C. Short Form Variants. A phrase that is a subphrase of a longer phrase, e.g. *change impacts - climate change impacts*.

The dictionary contains 5 507 unique phrases. Of these, 4 769 phrases have at least one synonym. Table 2 presents three example entries. The dictionary is organized alphabetically, and each phrase acts as a head phrase in its turn.

Table 2: Three example entries in the climate change RT-dictionary.

Head phrase	IS	Synonyms	IS
hydrologic cycle	11.3	hydrological cycle	15.1
ice melt	11.5	ice melting	5.7
		melting ice	11.8
		melting of ice	10.3
sea level rising	5.5	rising sea level	13.0
		rising sea levels	18.1
		sea level	26.2
		sea level rise	25.2
		sea level rises	10.8

5 EXPERIMENTS

In the experiments, we selected test phrases, formulated test queries based on the selected phrases, and run the queries in two search systems. This section describes these tasks. Section 6 presents the evaluation measures and reports the findings.

We investigated spelling and syntactic variants. As discussed above, short forms variants are often ambiguous expressions, and often it does not make sense to use them alone in queries. For this reason, they were not considered in these experiments. We may investigate them in the further research. They may be useful in document ranking, which is one possible application area of the RT-dictionary.

When the *test phrases* were selected from the dictionary it was ordered in the decreasing order of the ISs, so that high IS phrases were in the beginning and low IS phrases at the end of the dictionary. For both variant types, 25 entries were selected from the beginning and 25 from the middle of the dictionary. An entry contains a head phrase and its synonym(s) (Table 2). In the selection, a candidate entry was discarded if it did not have synonyms, then the next one was tried until there were 50 entries both for spelling and syntactic variants. The synonyms of the selected head phrases varied from high IS to low IS phrases. In the spelling variant test set, 38 head phrases had one synonym, 10 had two, one had three, and one had four synonyms. In the syntactic variant set, these numbers were: 36 head phrases had one synonym, 10 had two synonyms, and four had three synonyms.

Basically, we formulated 50 *test topics* both for spelling and syntactic variants. The test topics were of the type *climate change AND C*, where C is the concept represented by the entry selected for the experiment. Unlike in traditional information retrieval experiments (see http://trec.nist.gov), the relevance of the documents was not assessed by human assessors. Instead, we performed (1) *high precision* queries, that is, *title* queries, and *high recall* queries, that is, *full text* queries. In (1) the query phrase is required to appear in the title of the document and in (2) the query phrase may appear anywhere in the document. These two types represent a broad spectrum of search results regarding the precision and recall of the results. It should be noted that we are not interested in the precision and recall as such, but the effectiveness of the IS to indicate good query phrases in different situations.

If the phrase occurs in the title of the document, we can be highly confident that the document deals with the issue represented by the phrase. The title queries do not retrieve all documents that are relevant to a particular test topic. However, they retrieve highly relevant documents which usually are the most important for the user. The title queries are important also in that they return a focused set of relevant documents. Common experience suggests that Web searchers usually only look at the top search results. This has also been verified experimentally (Jansen et al., 2000). Thus, a relatively small set of relevant documents is more important than high recall. The full text queries return highly relevant, marginally relevant and irrelevant documents. They thus yield higher recall but lower precision than title queries.

We run the queries in our climate change search system and in the Google search engine. Google is the biggest search engine on the Web, covering billions of pages, and it is characterized by the heterogeneous information. Google‘s advanced search mode supports title queries. The climate change search system has indexed 95 819 (public version 73 194) Web pages related to climate change. The pages crawled for the system were indexed using the Apache Lucene programming library (http://lucene.apache.org/). The system supports several types of queries based on Lucene’s query language, e.g. queries targeted at the title of documents, and it reports the number of retrieved pages.

In the experiments, each test phrase was run as a query in the climate change search system (in short *CCSS* in tables) and in Google. All documents indexed by our search system deal with climate change and it was not explicitly expressed in queries, whereas Google queries included the phrase *climate change* in addition to the test phrases.

Google reports the number of retrieved pages which, however, may vary in that the same query returns different number of pages even during a short period of time. However, generally the retrieval results are consistent and meet the expectations.

Below we present some example title queries:

Climate change system - the highest IS phrase

title: "biodiversity loss" [IS=12.6]

Climate change system - other phrase

title: "loss of biodiversity" [IS=11.2]

Google - the highest IS phrase

allintitle: "biodiversity loss" "climate change"

Google - other phrase

allintitle: "loss of biodiversity " "climate change"

6 FINDINGS

In the first experiment, we compared the number of documents retrieved by the highest IS phrases to that retrieved by the other phrases. If the IS is a good indicator of query performance we expect the highest IS phrases to retrieve much larger sets of documents than the other phrases. Tables 3 and 4 report the total number of documents retrieved by the highest IS phrases and the other phrases across the synonym sets for title queries (Table 3) and for full text queries (Table 4). As can be seen, in both cases the highest IS phrases perform far better. The statistical significance was analyzed by the paired t-test. By conventional criteria, the results are statistically significant except for Google / syntactic / title. In Tables 3-4 the significance levels are indicated as follows: *** $p < 0.001$; ** $p < 0.005$; * $p < 0.05$.

The second experiment considered effective phrases in general. The test phrases were put in six categories of equal size, i.e., the same number of phrases, based on their ISs. For spelling variants, each category contained 19 phrases, and for syntactic variants 20 phrases. Tables 5 (title queries) and 6 (full text queries) report the average number of re trieved documents in each category, and the average IS for the categories. As can be seen, the trends are not strictly linear but exhibit downward curvatures in a few cases. However, the trends are still clear: The higher the IS, the more documents are retrieved. Linear regression analysis yielded the correlation coefficients from 0.87 (Google / spelling / title, $p < 0.05$) to 0.92 (CCSS / spelling / title, $p < 0.01$).

These indicate a strong relationship between the IS and the search results.

Table 3: Number of retrieved documents by the highest IS phrases and other phrases. Title queries.

Variant type *Search system*	*Highest IS* *# docs*	*Other* *# docs*
Spelling, CCSS	3545 ***	793
Syntactic. CCSS	1163 ***	295
Spelling, Google	686 854 *	331 556
Syntactic. Google	1 374 087	748 841

Table 4: Number of retrieved documents by the highest IS phrases and other phrases. Full text queries.

Variant type *Search system*	*Highest IS* *# docs*	*Other* *# docs*
Spelling, CCSS	53 839 ***	16 403
Syntactic. CCSS	49 146 ***	14 205
Spelling, Google	408 524 000 ***	133 057 300
Syntactic. Google	223 705 000 **	94 023 730

Table 5: Average number of retrieved documents in the six IS categories. Title queries.

Variant type. *Average IS for the category*	*CCSS* *# documents*	*Google* *# documents*
Spelling		
IS=24.3	134.3	22 453.5
IS= 20.1	76.9	26 923.3
IS= 15.1	6.7	3 325.8
IS=13.4	2.8	523.5
IS=11.9	6.8	296.2
IS=7.8	0.8	78.2
Syntactic		
IS=21.8	48.4	42 925.8
IS=17.6	17.2	28 877.6
IS=14.4	3.4	31 409.5
IS=12.8	4.7	2 419.1
IS=9.9	0.2	594.1
IS=6.9	0.2	1563.3

In summary, the results of the experiments showed that the importance score is an effective indicator of the effectiveness of the phrases in searching. The trends are clear:

- In a synonym set, the phrase with the highest IS retrieves significantly more documents than the other phrases.

Table 6: Average number of retrieved documents in the six IS categories. Full text queries.

Variant type. Average IS for the category	*CCSS # documents*	*Google # documents*
Spelling		
IS=24.3	1317,8	12 071 526.3
IS= 20.1	1612,1	11 427 842.1
IS= 15.1	452,5	1 460 737.8
IS=13.4	121,5	1 482 000.0
IS=11.9	128,6	1 457 136.8
IS=7.8	64,5	605 036.8
Syntactic		
IS=21.8	1983,4	7 620 500.0
IS=17.6	472,4	3 057 610.5
IS=14.4	347,4	1 434 650.0
IS=12.8	291,3	1 372 726.3
IS=9.9	73,4	1 265 878.9
IS=6.9	38,3	362 090.6

- The higher the IS is, the more documents are retrieved.
- The IS is effective in many search situations.

7 DISCUSSION

Web search engines are query-based search systems. Querying is an effective method when the information need can be expressed using a relatively simple query and the search term is clear, e.g. when the searcher looks for information on the disease whose name he or she knows. However, it is common that the searchers do not know or remember the appropriate search terms. Complex information needs are also difficult to formulate as a query. In situations such as complex work tasks the information need itself may be vague and ill defined (Ingwersen and Järvelin, 2005). Obviously, a terminology tool containing the most important phrases related to a particular topic would be a helpful tool for searchers searching for information related to that topic, helping to clarify the information need and find good query terms. We developed such a tool: a new type of dictionary that is intended as a search assistance in topic-specific Web searching. The dictionary has the following new features compared to standard dictionaries and thesauri: It contains real-text phrases and synonym groups, and each phrase is assigned an importance score.

Real-text Phrases. The phrases included in the real-text dictionary of climate change were extracted from the pages dealing with climate change, and are thus known to appear in the pages discussing climate change issues when used as search terms. Hence, the proposed approach implicitly involves the idea of reciprocity: the phrases are extracted from relevant Web pages (i.e., pages related to some aspect of climate change), and they in turn can be used in queries to find relevant pages.

Synonymy. It seems that Web searchers are not aware of the effects of synonymy on the search results, because not many search systems provide the searchers with terminology tools that support query formulation. The proposed RT-dictionary is such a tool that groups synonymous phrases together. The main limitation of our approach is that it does not find synonyms that are written completely differently. We intend to address this issue in the ongoing project. Wei et al. (2009) applied co-click analysis for synonym discovery for Web queries. The idea is to identify synonymous query terms from queries that share similar document click distribution. We cannot use the co-click analysis because we do not have sufficient amount of click data, but we can try a similar kind of idea: analyzing documents that share similar in- and out-link distribution.

Importance Score. The importance score was devised to be a measure that indicates how important the phrase is from the viewpoint of the search results. The results showed that it works as it was intended to work. It indicates effectively the best phrase among synonymous phrases and effective phrases in general. The IS is effective even in the Google search engine that has indexed billions of pages. Thus, the RT-dictionary provides the user with information on which of the alternative phrases is likely to yield the best search results. In cases where two (or more) query terms have high IS, the results of this study suggest that the user should use both (all) of them (provided that the search system supports disjunction (OR-operator) type queries).

RT-dictionary was developed for use as a search assistance to support query formulation in Web searching, but it could be used in document indexing as well to improve the ranking of documents. In Web search engines, pages are scored with respect to the input query and ranked on the basis of the assigned scores. Ranking schemes typically include a term frequency (tf) component, where term frequency refers to the number of occurrences of the query term in a document. The rationale behind the use of tf is that the more occurrences the query term has in a given document the more likely it is that the document is relevant to the input query. If synonymy is taken into account by summing up the term frequencies of synonyms in a document, more accurate relevance scores are achieved in comparison to a conventional approach where synonymy is not taken into account. This can be illustrated using a simple example. Consider a page containing the following phrases with each having one occurrence on the page: *sea level rise*, *rising sea level*, and *rising sea levels*. In the conventional situation where the user only uses base form query term (i.e., *sea level rise*) and the term frequencies of synonyms are not summed up tf=1, whereas when the term frequencies are summed up tf=3, which is more realistic because the concept denoted by the three phrases indeed appears three times on the page. Authors (of Web pages) tend to use alternative phrases and do not only use base form terms but also different syntactic and morphological forms. This means that many important documents are ranked lower than they actually deserve if synonymy is not taken into account.

8 CONCLUSIONS

We described a method to construct a topic-specific dictionary of real-text phrases to support query formulation in Web searching, and presented the existing climate change RT-dictionary. The proposed method is a general method and can be applied to any reasonable topic. We argued that there is need for such search assistances due to the difficulty to formulate queries in particular for complex information needs. In the experimental part of this paper we showed that the proposed importance score is a good indicator of search success.

The further development of RT-dictionary was discussed in the previous sections. Our plan is also to construct RT-dictionaries for new topics and to add multilingual features to the RT-dictionary.

ACKNOWLEDGEMENTS

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), 61-71.

Bergmark, D., Lagoze, C. and Sbityakov, A., 2002. Focused crawls, tunneling, and digital libraries. *Proc. of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, Rome, Italy, September 16-18, pp. 91-106.

Chakrabarti, S., van den Berg, M. and Dom, B., 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Proc. of the Eighth International World Wide Web Conference*, Toronto, Canada, May 11-14, pp. 1623-1640.

Cronen-Townsend, S., Zhou, Y. and Croft, B., 2002. Predicting query performance. *Proc. of the 28th ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, August 11-15, pp. 299-306.

Diligenti, M., Coetzee, F. M., Lawrence, S., Giles, C.L. and Gori, M., 2000. Focused crawling using context graphs. *Proc. of the 26th International Conference on Very Large Databases (VLDB)*, Cairo, Egypt, September 10-14, 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), 132-144.

He, B. and Ounis, I., 2006. Query performance prediction. *Information Systems*, 31(7), 585-594.

Ingwersen, P. and Järvelin, K., 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. Heidelberg, Springer.

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

Jansen, B. J., Spink, A. and Saracevic, T., 2000. Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing & Management*, 36(2), 207-227.

Lee, H. J., 2008. Mediated information retrieval in Web searching. *Proc. of the American Society for Information Science and Technology*, 45(1), pages 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), 892-910.

Perez-Iglesias, J. and Araujo. L., 2010. Standard deviation as a query hardness estimator. *The 17th International Symposium on String Processing and Information Retrieval (SPIRE 2010)*, Los Cabos, Mexico, October 11-13, pp. 207-212.

Pirkola, A., 2011a. Constructing topic-specific search keyphrase suggestion tools for Web information retrieval. *Proc. of the 12th International Symposium on Information Science (ISI 2011)*, Hildesheim, Germany, March 9-11, pp. 172-183.

Pirkola, A., 2011b. A Web search system focused on climate change. *Digital Proceedings, Earth Observa-*

tion of Global Changes (EOGC), Munich, Germany, 13-15 April.

Robertson and Willett., 1998. Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1), 48-69.

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

Tang, T., Hawking, D., Craswell, N. and Griffiths, K., 2005. Focused crawling for both topical relevance and quality of medical information. *Proc. of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)*, Bremen, Germany, October 31-November 5, pp. 147-154.

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

Wei, X., Peng, F., Tseng. H., Lu, Y. and Dumoulin, B., 2009. Context sensitive synonym discovery for web search queries. *Proc. of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*, Hong Kong, pp. 1585-1588.

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