Query expansion based on conceptual and contextual term relationships in Wikipedia

Wira-Alam, Andias; Saad, Farag; Mutschke, Peter

Query Expansion based on Conceptual and Contextual Term Relationships in Wikipedia

Andias Wira-Alam, Farag Saad, Peter Mutschke

GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8, D-50667 Cologne, Germany
{andias.wira-alam|farag.saad|peter.mutschke}@gesis.org

Abstract

The dramatic increase in information on the World Wide Web makes it more difficult for web users using web search engines to effectively satisfy their information needs. The users' lack of knowledge regarding the searched topics creates a complicated problem when formulating an effective query. Query expansion can play an essential role in overcoming such a deficit. However, because they lack sufficient knowledge about the searched topics, users sometimes find it difficult to evaluate the relatedness of the system's automatically expanded terms. This problem arises mostly in domain-specific areas, e.g. Social Science. In this paper, we expand the queries based on a structured, open knowledge resource on the Web (Wikipedia). We link the entities from a domain-specific corpus (qualitative journals in social science) to Wikipedia entities. With Wikipedia serving as background knowledge, we help users with their selections by providing the most likely related terms. Furthermore, users are provided with contextual information that describes each expanded term in order to give users a clearer idea about the meaning of each expanded term. By utilizing 10 test queries chosen by experts who also evaluated the results, we compared the results of using Wikipedia with the results achieved by using a qualitative journal.

In: H.-C. Hobohm (Hrsg.). Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten. Tagungsband des 13. Internationalen Symposiums für Informationswissenschaft (ISI 2013), Potsdam, 19.–22. März 2013. Glückstadt: Verlag Werner Hülsbusch, 324–338.

1 Introduction

The dramatic increase in information on the Web makes it more difficult for users to effectively satisfy their information needs. Web search engines are not effective if the information needs are not properly formed. This is especially true for casual users, whose web search usage has been growing exponentially. Due to lack of knowledge regarding the searched topics, this in turn leads to query terms not being matched to terms in the searched documents (vocabulary mismatch) (Custis/Al-Kofahi 2007; Furnas et al. 1987). Users cannot always formulate their queries properly when they are not familiar with the topics being searched. This increases the gap between the optimal query terms that should be used and the query terms actually used by the user.

Another issue is that web search queries are often short, generally between 2.4 and 2.7 words long (Gabrilovich et al. 2009), which does not provide enough context for effective information retrieval. For these types of queries, search engines do not usually provide high quality results. For example, search engines cannot properly handle homonyms (words that share the same spelling but have different meanings) e.g., Java the island or JAVA the programming language. Here, interactive query expansion (Fonseca et al. 2005) could play a pivotal role in disambiguating the user's query by adding new term/terms to the original query. For example, the query "java" can be ambiguous as it could have two different meanings, i.e. "Java programming language" or "Java Island".

Query expansion can be helpful by including relevant terms related to the original query, thus improving precision and recall. For example, when a user submits the query "car", documents containing "auto", "automobile", "sedan", "vehicle", etc. are not likely to be retrieved. These missing terms from the original query can affect precision and recall (Xu/Croft 2000). To alleviate the above mentioned issues and to improve the performance of a search engine, query expansion approaches have been widely used to support users and better satisfy their information needs (Cao et al. 2008; Chirita/Firan/Nejdl 2007; Mutschke et al. 2011). By using query expansion, new terms can be added to the original query so more documents can be retrieved.

In this paper, we expand the user query with related terms from Wikipedia. Our approach is to exploit the hyperlink structure in Wikipedia in order to determine closely related concepts based on linking entities and the con-

tent of related pages. Query expansion is then done based on concepts rather than with pure words. For instance, when submitting a search for the German sociologist "Jürgen Habermas", the query would be much more effective when expanded by the concept "Critical Theory" which is strongly related to "Jürgen Habermas". Based on this expanded concept, further concepts could be added, such as "Theodor W. Adorno", another German sociologist related to "Jürgen Habermas", and so on. This information is provided using knowledge extracted from Wikipedia, a significant body of information that is constructed and evaluated by humans in a conceptual relationship-based form.

The paper is structured as follows: in the next section we describe the proposed approach in detail. In Section 3, we show the evaluation of the results in details. In Section 4, we review other research efforts related to our work. The last section contains concluding remarks and reflections on the direction of future work.

2 The Proposed Approach

Terms in Wikipedia are represented by articles. An article is an entity that describes a concept of class, person, place or other subject. Wikipedia aims to make resources identifiable on the Web to ensure availability and accessibility for web users. Terms that are directly connected to each other can be recognized as related. As an illustration, we see in Figure 1 that the article "Jürgen Habermas" in Wikipedia is connected to some other articles by its outlinks and backlinks, such as "Kritische Theorie". Intuitively, we might consider that the concepts representing these articles are somehow related and have coherence with "Jürgen Habermas".

Based on the aforementioned illustration, we make use of the hyperlink structure in Wikipedia in order to extract related terms. However, these terms need to be further selected for expanding user queries in domain-specific information retrieval. In a scientific corpus, such a link structure can be obtained e.g. by extracting the citation graph. However, the citation graph does not express the term relationship, but it does show the connection between documents.

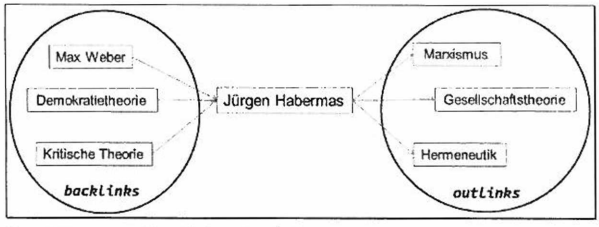

Fig. 1 Illustration of backlinks and outlinks

As a matter of fact, in a domain-specific search, users are mainly interested in searching for concepts or author names, which are actually entities. An article in Wikipedia generally contains a brief description as well as comprehensive information about a particular entity and its relation to other entities. In contrast, a document in a scientific corpus does not employ relationships between entities but instead contains a raw description about a particular topic. Compared to Wikipedia, query expansion using a corpus of scientific documents would therefore miss important background information such as authors' biography or relationships between authors and institutions. As seen in Figure 2, it is possible to enrich the document corpus with e.g. biographical information taken from Wikipedia. Therefore, our approach is to complement the information contained in a corpus of scientific documents using background information taken from Wikipedia. For this we compute the co-occurrence between concepts in Wikipedia and expand the user's query with the concepts from Wikipedia that are most closely related to the query terms matching the concept.

Due to the availability of training data and experts to evaluate the proposed approach, we decided to focus on the Social Science domain. We used two training corpora for this work. The first corpus is the German Wikipedia, the second corpus, as a reference corpus, is a set of qualitative journals provided by the German Sociological Association (GSA) consisting of about 6,000 documents.

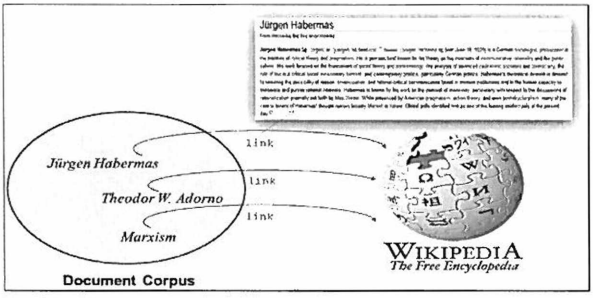

Fig. 2 Enriching document corpora using linked entities from Wikipedia

Wikipedia Corpus

In the German Wikipedia there are about 2 million articles. In order to reduce the complexity, we first of all extracted a particular collection of German Wikipedia articles that are related to the Social Sciences. We used the Thesaurus for Social Sciences (TheSoz)[1] to determine this subset. TheSoz is a crucial instrument for content-oriented searches by keywords, containing about 12,000 entries and covering topics in all of the Social Science disciplines. Additionally, we also used a list of notable authors in the GSA corpus, the most part of which is represented by a Wikipedia page.[2] This may be helpful later on for query expansions in special cases when focusing on expanding information related to author names.

Articles Collection

As stated above, we have a list of TheSoz terms and authors. The important task now is to link these entities to Wikipedia. For this purpose, we downloaded the list of page titles provided by Wikimedia database dumps[3] and matched Wikipedia entries against the list of authors and TheSoz terms.

1 http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus <February 19, 2013>

2 The list of authors can be extracted from the metadata fields that are maintained in the Database SozDB at GESIS, in total there are about 3,800 authors.

3 http://dumps.wikimedia.org/dewiki/20120603/dewiki-20120603-all-titles-in-ns0.gz <February 19, 2013>

Thus, we can formally describe the task of finding the relevant subset T from all Wikipedia articles as follows

$$T = W \cap (X \cup A)$$

where T is a subset of Wikipedia entries or titles W that are matched with the TheSoz terms X and authors A. This task is implemented by using string matching with some preprocessing (see Table 1 as an example). Finally, after we obtained T, we crawled each article page $\in T$, extracted the page contents, stripped the HTML tags, and stored them into individual text files for the training corpus that had been indexed. As a result, we obtained 8,270 matches for the terms[4] and 963 for the authors.

Table 1. TheSoz terms and authors matched to Wikipedia titles

TheSoz Terms / Authors	Wikipedia Titles
Habermas, Jürgen	Jürgen_Habermas
Frankfurter Schule	Frankfurter_Schule
empirische Sozialforschung	Empirische_Sozialforschung

Expanding Query Terms with Wikipedia Collection

Since each term is represented as an article, given a query term, q, the related terms are the articles that are most similar to the source. First of all, we match q with our collections T and W in order to find the corresponding document. To be precise, if there is only a match in W, we add q into T and update the index. Afterwards, we extract another subset from its backlinks and outlinks by using a method described by Wira-Alam (2012), denoted as $L^q_{\leftarrow}$ and $L^q_{\rightarrow}$, respectively[5].

After obtaining $L^q_{\leftarrow}$ and $L^q_{\rightarrow}$, we match them against T and, if there are relative complements $L^q_{\leftarrow} \cap T^C$ or $L^q_{\rightarrow} \cap T^C$, add those elements into T. Finally, the query expansion task can be seen as 'finding related documents'. Formally, we define the similarity score of two documents, denoted as d_1 and d_2, as follows

4 These matches have been added as a new property (skos exactMatch) at the SPARQL endpoint of TheSoz and can be accessed online at http://lod.gesis.org/pubby/page/thesoz/ <February 19, 2013>

5 We could also use DBPedia since outlinks are also provided, but extracting backlinks would be a non-trivial task.

$$sim(d_1, d_2) = \frac{1}{n}|d_1 \cap d_2|$$

where both d_1 and d_2 are considered vectors of words where each has cardinality of n. The factor n is used to denote the number of important words included in the calculation.[6] The *importance* of a word can be obtained by calculating its *tf-idf* score and this method is a slight modification of *Jaccard similarity coefficient*. As stated before, we prioritize the terms that are contained in $L^q_{\leftarrow}$ and $L^q_{\rightarrow}$ to be shown in the results list. However, we set a threshold value of 0.8 for the expanded terms in order to avoid vagueness.[7]

GSA Corpus

This corpus consists of about 6,000 documents and each document is a text file. By the same procedure as before, all documents had been indexed. Given a query term, the task is to find the top 10 related terms. Unlike finding matching documents using *tf-idf*, the score is calculated as follows:

$$P(t_2 \mid t_1) = \frac{freq(t_1, t_2)}{freq(t_1)}$$

where $freq(t_1,t_2)$ is the frequency of co-occurrence between t_1 and t_2, while $freq(t_1)$ is the number of documents containing t_1 in the training corpus. This method is based on *maximum likelihood estimate* (MLE).[8] This method is currently implemented in the state-of-the-art application for internal use.

6 In order to reduce the complexity, *stopwords* are ignored and n was set to 25. Furthermore, we also plan to implement a cache to prevent a repetition especially in the extracting and indexing process.

7 This value is not a strict value and can be adjusted e.g. if the number of results is very limited.

8 Note that *stopwords* are also ignored in order to increase the performance.

3 Evaluation

3.1 Journal Articles vs. Wikipedia

In the evaluation, we were interested in evaluating the accuracy of the proposed method. Furthermore, we also evaluate whether the contextual information provided by the algorithm is appropriate enough to guide the user in having wide understanding of the expanded terms. For the evaluation we picked 10 common terms chosen by two experts. These terms were chosen because they have a wide spectrum of discussions in the field. Moreover, the terms would also be interesting for the domain experts to evaluate the extracted contextual information in order to see the text quality of the text extracted from the Wikipedia. Thus, from each term, we extracted 10 related terms from each training corpus.

As seen in Table 2, the results extracted from the Wikipedia corpus are generally much better. Since Wikipedia is a well-structured knowledge base, Wikipedia provides better results. The results provided by the GSA corpus are at some points good as well. However, since the data is quite unstructured, e.g. containing no named entities, the results show some biases, especially for the multi-word terms.

Another important aspect to be considered is the quality of the contents in both corpora. Wikipedia articles are written primarily for those with little background knowledge, while the GSA corpus is a set of highly qualitative articles dedicated for use by experts. However, in order to provide good results, the quality of the contents in the corpora plays an unimportant role. Nevertheless, one cannot simply rely on Wikipedia, writing articles on Wikipedia is voluntary. The results of the query term "Chicago-Schule" are shown as being completely irrelevant. The problem lies in the length of the article and the extensive description in the text about other related articles.

Thus, we can prove our claim based on the achieved results that Wikipedia is a high quality resource which can be used without any further processing. Other available resources such as the GSA corpus are poor quality resources for query expansion as they need further processing to obtain high quality terms.

Table 2: Evaluation of the results: x(sign)

Terms	Expanded Terms (DGS)	Expanded Terms (Wikipedia)	Expert 1	Expert 2
Dialektik	horkheimer, kritischen, frankfurter, theorie, adorno, marcuse, subjektivitat, schule, horkheimers, kritische	Dialektischer Materialismus, Marxismus, Marxistische Theorie, Materialistische Geschichtsauffassung, Historischer Materialismus, Totalität, Idealismus, Argumentation, Politische Theorie, Metaphysik	3(+), 2(0), 5(?) / 6(+), 4(0)	3(+), 2(0), 5(?) / 8(+), 2(0)
Hermeneutik	kultursoziologischer, sprache, kultur, musik, musiksoziologie, symbolische, verstehens, kommunikation, sozialforschung, rechts	Verstehen, Medienanalyse, Geisteswissenschaftliche Pädagogik, Geisteswissenschaft, Geschichtsbewusstsein, Erkenntnis, Geschichtsphilosophie, Phänomenologie, Literaturinterpretation	2(+), 3(0), 4(?), 1(-) / 1(+), 4(0), 4(?), 1(-)	4(+), 3(0), 3(-) / 4(+), 4(0), 2(-)
Marxismus	arbeiterbewegung, krise, kommunistischen, parteien, kapitalistischen, arbeiterklasse, marxistischen, sozialdemokratie, bewegung, revolutionaren	Marxistische Theorie, Marxismus-Leninismus, Leninismus, Kommunismus, Weltrevolution, Marxistische Soziologie, Gesellschaftssystem, Gesellschaftsformation, Diktatur des Proletariats, Wissenschaftlicher Sozialismus	3(+), 1(0), 4(?), 2(-) / 3(+), 2(0), 4(?), 1(-)	10(+) / 10(+)
Materialismus	historischen, theorie, wirtschaftswissenschaften, werthaltungen, freizeit, gesellschaftsformationen, strukturen, interdependenz, organisationsprinzipien, produktionsweise	Dialektischer Materialismus, Historischer Materialismus, Materialistische Geschichtsauffassung, Idealismus, Deutscher Idealismus, Realität, Metaphysik, Naturphilosophie, Positivismus, Leben	1(+), 3(0), 1(?), 5(-) / 1(+), 3(0), 1(?), 5(-)	7(+), 2(0), 1(-) / 4(+), 3(0), 3(-)
Moderne	ungleichheit, foucault, herrschaft, rosa, modernisierung, uneindeutigkeit, grenzziehungen, nassehi, sozialer, reflexiver	Kunst, Expressionismus, Avantgarde, Okkultismus, Bildende Kunst, Dadaismus, Naturphilosophie, Tradition, Kunstmaler, Kulturpessimismus	4(0), 2(?), 4(-) / 2(+), 5(0), 2(?), 1(-)	3(0), 7(-) / 5(0), 5(-)

Terms	Expanded Terms (DGS)	Expanded Terms (Wikipedia)	Expert 1	Expert 2
Konstruktivismus	luhmann, umwelt-soziologie, realität, systemtheorie, umwelt, luhmanns, soziologie, form, gesellschaft, logie	Soziale Konstruktion, Wissenschaftlichkeit, Wissenschaftstheorie, Expressionismus, Metatheorie, Realität, Lernpsychologie, Relativismus, Architektur, Sozialistischer Realismus	2(0), 2(?), 6(-) / 3(0), 2(?), 5(-)	3(+), 2(0), 5(?) / 4(+), 2(0), 4(-)
Chicago-Schule	berufsfindung, wundt, leipziger, arbeitsmarkt, jugendlichen, natur, lamprecht, elfriede, bremen, soziologie	Fiskalismus, Keynesianismus, Österreichische Schule, Schule, Neoliberalismus, Schulbesuch, Schulbildung, Weiterführende Schule, Hans Dietrich Schultz, Frankfurter Schule	1(0), 4(?), 5(-) / 1(0), 3(?), 6(-)	10(-) / 10(-)
Frankfurter Schule	horkheimer, marx, kritischen, adorno, subjektivitat, kritik, subjekt, dialektik, frankfurt, objektiv	Kritische Theorie, Max Horkheimer, Theodor W Adorno, Herbert Marcuse, Negative Dialektik, Kulturindustrie, Dialektik, Marxistische Soziologie, Populärkultur, Popkultur	6(+), 2(0), 1(?), 1(-) / 6(+), 3(0), 1(?)	9(+), 1(0) / 9(+), 1(0)
Kritische Theorie	normalarbeitsverhältnisses, normalarbeitsverhältnis, rationalität, reellen, revolutionstheorie, lebendigen, rethel, arbeit, vernunft, prozeanalyse	Max Horkheimer, Frankfurter Schule, Theodor W Adorno, Herbert Marcuse, Negative Dialektik, Politische Theorie, Dialektik, Positivismusstreit, Ideologiekritik, Ideologie	2(+), 2(0), 2(?), 4(-) / 3(+), 1(0), 2(?), 4(-)	7(+), 3(0) / 9(+), 1(0)
Grounded Theory	latour, latours, vergleichbarkeit, vorwissen, forschung, verfahren, prozeduralen, oevermann, forschungsprozess, lorenz	Gerhard Kleining, Qualitative Methode, Qualitative Forschung, Feldforschung, Sozialforschung, Empirische Sozialforschung, Quantitative Methode, Soziografie, Soziologie, Soziologische Theorie	2(+), 2(0), 3(?), 3(-) / 3(+), 2(0), 2(?), 3(-)	4(+), 3(0), 3(-) / 5(+), 3(0), 2(-)

x denotes the number of terms, (+) means highly relevant, (0) moderately relevant, (?) unclear/ambiguous, (-) irrelevant

Furthermore, improving such resources is time consuming with no guarantee of obtaining high quality terms using Named Entity Recognition (NER). In other words, Wikipedia is a high quality structured resource which is continually growing, while other resources such as textual corpora require extensive effort in order to be improved.

3.2 Use of Context

The second experiment we conducted was to evaluate the usefulness of the contextual information for 10 randomly selected term pairs (the source term with its top ranked expanded term). The evaluation was conducted by experts who were requested to give a score between 0 (low) and 5 (high). We provided three examples of contextual information for each term pair. Experts had to evaluate each contextual information example individually and rate it with a score. Table 3 shows two examples of the contextual information evaluation.

Table 3:
Two examples of the obtained contextual information for two term pairs

Term Pairs	Extracted Contexts	Average Score
Marxismus and Marxistische Theorie	Als Kritische Theorie wird eine von Hegel, Marx und Freud inspirierte Gesellschaftstheorie bezeichnet, deren Vertreter auch unter dem Begriff Frankfurter Schule zusammengefasst werden. Ihr Gegenstand ist die kritische Analyse der bürgerlich-kapitalistischen Gesellschaft, das heißt: die Aufdeckung ihrer Herrschafts- und Unterdrückungsmechanismen und die Entlarvung ihrer Ideologien, mit dem Ziel einer vernünftigen Gesellschaft mündiger Menschen.	4
Konstruktivismus and Soziale Konstruktion	Hauptvertreter: Ernan McMullin, Stathis Psillos, ihrem Selbstverständnis nach auch Hilary Putnam und Richard Boyd, obwohl Putnams interner Realismus und Boyds Konstruktivismus bezüglich natürlicher Arten etwas von den klassischen Doktrinen abweicht.	1

These were performed by the algorithm (Mathiak 2012) showing one with a good result and another with a bad result.[9] We have slightly modified the original application in order to support German Wikipedia. Please try it yourself, e.g. with a term pair: "Frankfurter Schule" and "Kritische Theorie", as shown in Table 3.

Ultimately, in some cases the algorithm provided useful contextual information and achieved an overall average of 2.65 out of 5. However, in other cases the algorithm provided less useful contextual information. The reason for this deficit in providing useful contextual information can be related to the source of the contextual information (Wikipedia). For some term pairs, we could not find optimal contextual information that describes their relationship in the most advantageous way. In order to tackle such a deficit in the future, our plan is to work on enlarging our source data to obtain contextual information from other corpora.

4 Related Work

Query expansion approaches can be classified into two main approaches, namely global and local analysis (Xu 1996). In the global analysis approach, word occurrences and relationships in the corpus as a whole are examined. Based on this examination, a list of candidate words are extracted and added to the original query. Since global analysis examines the corpus as a whole, it is assumed to be less efficient for performing the expansion task when compared to the local analysis approach. The local analysis approach examines only the top ranked documents that are obtained using the original user query. These top-ranked documents can be obtained by the system automatically (blind-feedback) (Xu/Croft 1996) or by user feedback (pseudo-feedback) (Salton/Buckley 1997) where a user can judge which documents may or may not be relevant to the given query. However, some of the top-retrieved documents might not be exactly relevant to the user's information needs.

9 The excerpts are extracted using http://multiweb.gesis.org/RelationShipExtractor_DE/ <February 19, 2013>

Other query expansion approaches make use of structured knowledge resources such as WordNet, Web or Wikipedia. For example, Yang et al. (2003) made use of WordNet and Web in order to expand the user query and applied it to the Question Answer (QA) problem. Wikipedia has also been used to expand the user query, while Arguello et al. (2008) made use of links and anchor text in Wikipedia in order to expand the user query.

Another example of using Wikipedia as a high quality resource for query expansion is proposed in Müller/Gurevych (2008). The authors evaluated the performance of several information retrieval models by using the content of Wikipedia articles. In this paper, however, we expand the queries by leveraging the thesaurus terms in order to obtain high quality expanded terms.

Recently, our approach has been used to expand queries for the CHiC 2012 pilot evaluation and obtained promising results (Schaer et al. 2012). However, no context extraction was performed. To the best of our knowledge, there was no attempt at combining the hyperlink structure within Wikipedia and context extraction to support query expansion for domain-specific information retrieval.

5 Conclusion and Future Work

We showed in this paper that the Web is a suitable resource for the background knowledge needed for query expansion in a domain-specific field. We believe that this work contributes to the development of information retrieval in general. By leveraging simple methods, we achieved promising results in a real use case. Wikipedia is continually growing, quantitatively and qualitatively, therefore our approach offers a number of benefits and a variety of uses. Since we also have a list of authors, we will also offer this feature to users in the future, e.g. if users desire only author names to be shown. However, our approach also has limitations since we currently only consider terms represented in Wikipedia and there is no suggestion for term disambiguation.

Nevertheless, as a proof-of-concept, in this paper we have also evaluated the use of contexts to help users gain a better understanding between the original and expanded query terms, not only targeting domain experts but also young scientists. As mentioned previously, this investigation was trig-

gered by the question: "How should users decide which expanded terms are to be chosen for further search?" The current results show that contextual information about the expanded terms, while preliminary, suggests a promising direction for further investigation.

In the future, we therefore plan to build an interactive user interface, as seen in Figure 3, so that users can select or deselect the related terms based on the given excerpts. Nevertheless, the quality of the content now plays an important role and therefore it is a big challenge for us to work on improving the current results. Moreover, we also plan to further investigate our approach with other domain-specific document corpora for a deeper evaluation, e.g. to perform an extrinsic task.

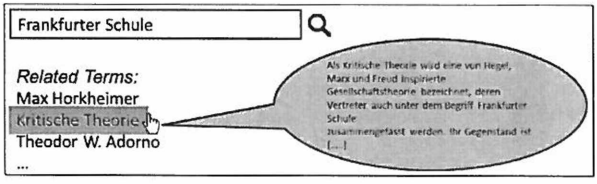

Fig. 3 A mockup screenshot for the planned interactive user interface in the future

Acknowledgements

We would like to thank Andreas Kempf and Reinhard Messerschmidt for their work on the evaluation and for their useful input and advice on improving our work.

References

Arguello, J.; Elsas, J. L.; Callan, J.; Carbonell, J. G. (2008). Document representation and query expansion models for blog recommendation. In: 2nd Intl. Conf. on Weblogs and Social Media (ICWSM) 2008.

Cao, G.; Nie, J. Y.; Gao, J.; Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. In: ACM SIGIR'08, 243–250.

Custis, T., Al-Kofahi, K. (2007). A new approach for evaluating query expansion: query document term mismatch. In: ACM SIGIR'07, 575–582.

Chirita, P. A.; Firan, C. S.; Nejdl, W. (2007). Personalized query expansion for the web. In: ACM SIGIR'07, 7–14.

Fonseca, B. M.; Golgher, P.; Pôssas, B.; Ribeiro-Neto, B.; Ziviani, N. (2005). Concept based interactive query expansion. In: ACM CIKM '05, 696–703.

Furnas, G. W., Landauer, T. K., Gomez, L. M., Dumais, S. T. (1987). The vocabulary problem in human-system communication. In: Commun. ACM 30 (11), 964–971. http://doi.acm.org/10.1145/32206.32212 <February 19, 2013>.

Gabrilovich, E.; Broder, A.; Fontoura, M.; Joshi, A.; Josifovski, V.; Riedel, L.; Zhang, T. (2009). Classifying search queries using the web as a source of knowledge. In: ACM Transactions on the Web 3 (2), 1–28.

Mathiak, B.; Martínez-Peña, V. M.; Wira-Alam, A. (2012). What is the relationship about? extracting information about relationships from wikipedia. In: WEBIST 2012, 625–632.

Müller, C., Gurevych, I. (2008). Using wikipedia and wiktionary in domain-specific information retrieval. In: CLEF'08. Heidelberg: Springer, 219–226.

Mutschke, P.; Mayr, P., Schaer, P.; Sure, Y. (2011). Science models as value-added services for scholarly information systems. In: Scientometrics 89, 349–364. http://dx.doi.org/10.1007/s11192-011-0430-x <February 19, 2013>.

Salton, G.; Buckley, C. (1997). Improving retrieval performance by relevance feedback. In: Sparck Jones, K.; Willett, P. (Hrsg.) Readings in information retrieval. San Francisco: Morgan Kaufmann, 355–364.

Schaer, P.; Hienert, D.; Sawitzki, F.; Wira-Alam, A.; Lueke, T. (2012). Dealing with sparse document and topic representations. In: CLEF/CHiC Workshop-Notes. http://arxiv.org/abs/1208.3952 <February 19, 2013>.

Wira-Alam, A.; Mathiak, B. (2012). Mining wikipedia's snippets graph – first step to build a new knowledge base. In: Völker, J.; Paulheim, H.; Lehmann, J.; Niepert, M. (Hrsg.). Proceedings of the First International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data. urn:nbn:de:0074-868-4, 43–48.

Xu, J.; Croft, W. B. (1996). Query expansion using local and global document analysis. In: ACM SIGIR'96, 4–11.

Xu, J.; Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. In: ACM Trans. Inf. Syst. 18 (1), 79–112.

Yang, H.; Chua, T. S.; Wang, S.; Koh, C. K. (2003). Structured use of external knowledge for event-based open domain question answering. In: ACM SIGIR'03, 33–40. http://doi.acm.org/10.1145/860435.860444 <February 19, 2013>.