HITS and Misses: Combining BM25 with HITS for Expert Search

Johannes Leveling and Gareth J. F. Jones

School of Computing and Centre for Next Generation Localisation (CNGL)
Dublin City University
Dublin 9, Ireland
{jleveling, gjones}@computing.dcu.ie

Abstract. This paper describes the participation of Dublin City University in the CriES (Cross-Lingual Expert Search) pilot challenge. To realize expert search, we combine traditional information retrieval (IR) using the BM25 model with reranking of results using the HITS algorithm. The experiments were performed on two indexes, one containing all questions and one containing all answers. Two runs were submitted. The first one contains the combination of results from IR on the questions with authority values from HITS; the second contains the reranked results from IR on answers with authority values. To investigate the impact of multilinguality, additional experiments were conducted on the English topic subset and on all topics translated into English with Google Translate. The overall performance is moderate and leaves much room for improvement. However, reranking results with authority values from HITS typically improved results and more than doubled the number of relevant and retrieved results and precision at 10 documents in many experiments.

Key words: Expert Search, Information Retrieval, BM25, HITS Algorithm

1 Introduction

The CriES pilot challenge [1] aims at multilingual expert search and is based on a subset of the data provided by Yahoo! Research Webscope[1]. The complete Yahoo QA dataset comprises 4.5M natural language questions and 35.9M answers. Questions are associated with one or more answers and the best answer is marked by users of the web portal. Questions are also annotated with categories from a hierarchical classification system. The Yahoo QA dataset has been previously used in [2] to train a learning to rank approach. The CriES data subset was extracted with the preprocessing tool provided by the organizers. This subset contains 780,193 questions, posted by more than 150,000 users.

For the CriES expert search experiments described in this paper, different approaches to find experts likely to answer a question were investigated: 1. Finding experts by matching the current question with previously given answers.

---

[1] http://research.yahoo.com/

This corresponds to a standard information retrieval approach on answer documents. 2. Finding experts by matching the current question with questions which have previously been answered. This approach is typically employed in FAQ (frequently asked questions) search and corresponds to IR on questions. 3. + 4. Reranking the results of the two former approaches by interpreting HITS authority values of question and answer documents as the level of expertise.

The rest of this paper is organized as follows: Section 2 introduces related work. Section 3 describes the theoretical background and the system setup for the CriES experiments. Section 4 presents the experimental setup and results, followed by an analysis and discussion of results in Section 5 and the paper concludes with an outlook on future work in Section 6.

2 Related Work

Expert search on question answer (Q/A) pairs is a relatively new research area which is related to search in FAQs, social network analysis, and question answering (see, for example [3]).

2.1 FAQ Search

Burke, et al. [4, 5] introduce FAQ finder, a system for finding answers to frequently asked questions. Their experiments are based on a small set of FAQ files from Usenet newsgroups. A weighted sum of vector similarity between question and Q/A pairs, term overlap, and WordNet-based lexical similarity between questions is computed to find the best results.

Wu, et al. [6] use a probabilistic mixture model for FAQ finding in the medical domain. Questions and answers are first categorized and the Q/A pairs are interpreted as a set of independent aspects. WordNet [7] and HowNet [8] are employed as lexical resources for question classification. Answers are paragraphed, and clustered by LSA and k-means. A probabilistic mixture model is used to interpret questions and answers based on independent aspects. Optimal weights in the probabilistic mixture model are estimated by expectation maximization. This approach outperforms FAQ finder [4] in the medical domain.

Jijkoun and de Rijke [9] describe FAQ finding based on a collection of crawled web pages. Q/A pairs are extracted and questions are answered by retrieving matching pairs. The approach is based on the vector space model and Lucene, using a linear combination of retrieval in different fields.

Chiu, et al. [10] use a combination of hierarchical agglomerative clustering (HAC) and rough set theory for FAQ finding. HAC is applied to create a concept hierarchy. Lower/upper approximation from rough set theory helps to classify and match user queries. They conclude that rough set theory can significantly improve classification of user queries.

Several retrieval experiments described in this paper are also based on finding experts who answered similar questions by indexing all questions.

2.2 Expert Search

Balog, Azzopardi, and de Rijke [11] propose two models to find experts based on documents for the TREC enterprise track[2]. The first approach is to locate knowledge from experts' documents; the second approach aims at finding documents on topics and extract associated experts. To this end, they analyze the communication link structure. They find that the second approach consistently outperforms the first one.

MacDonald and Ounis [12] perform experiments on expert finding on the TREC enterprise data. They find that increasing the precision in the document retrieval step does not always result in better precision for the expert search.

Yang, et al. [13] present the expert finding system EFS, which employs experts' profiles created from their lists of publications. Category links are extracted from Wikipedia. Nine different areas of expertise are differentiated.

Similar to extracting experts from retrieved documents, some retrieval experiments described in this paper rely on retrieving answers to a given question and extracting their experts (authors).

2.3 Link Analysis

The HITS (Hyperlink-Induced Topic Search) algorithm is a link analysis algorithm for rating web pages [14]. PageRank [15] produces a static, query-independent score for web pages, taking the incoming and outgoing links of a web page into account. In contrast, HITS produces two values for a web page: its authority and its hub value. HITS values are computed at query time and on results retrieved with an initial retrieval, i.e. the computations are performed only on initially retrieved results, not across all linked web pages. Recent variants of HITS have been concerned with stability of the algorithm [16] and with modifications of the algorithm to improve precision [17].

For our CriES experiments, we selected the HITS algorithm, as its values are computed at query time and on a smaller document base. Thus, the HITS algorithm does not require re-indexing the document collection to recompute scores after modifications or extensions to the algorithm. HITS scores also highly correlate with in/outdegree of linked nodes, which intuitively correspond to the level of expertise: the more information a person produces on a given topic, the higher her/his level of expertise should be.

The experiments for the CriES pilot challenge can be based on two different types of data which were provided by the organizers: a collection of Q/A pairs and a linked graph model extracted from this collection. Furthermore, the CriES challenge is unique in that it aims at expert finding in a multilingual setting, i.e. topics are provided in different languages.

---

[2] http://www.ins.cwi.nl/projects/trec-ent/

3 System Description

3.1 Topic and Document Processing

Interpreting individual questions and answers as documents, standard IR techniques can be applied for expert search. In our work, the Lucene toolkit[3] was utilized to preprocess the topics and documents, and to index and search the document collection. Standard Lucene modules were employed to tokenize the questions and answers and to fold upper case characters to lower case. Stopword lists from Jacques Savoy's web page on multilingual IR resources[4] were used to identify stopwords. Stemming of topics and documents was performed using the Snowball stemmer for the corresponding language provided in Lucene. For all retrieval experiments, only the topic fields for *'title'* and *'description'* were used to create IR queries for Lucene (TD). The fields *'narrative'* and *'questioner'* were omitted for query formulation. The *'answerer'* field was used to form documents IDs. Figure 1 shows a sample topic.

The CriES question-answer set was preprocessed by us to generate two types of documents from the original CriES documents: answer documents (A) and question documents (Q). The first type of document contains the *'answerer'* ID as a document ID and the text of his answer concatenated with the category of the question. This retrieval approach realizes standard IR by finding answers based on the replies the users have already generated. The second type of document contains the *'answerer'* ID as a document ID and the question text concatenated with all category labels from the original document. Thus, retrieval on these documents aims at finding experts by matching the input question with previous questions the answerer has replied to. In detail, documents for indexing were created as follows: answer documents were extracted from answers given (i.e. *'bestanswer'*); question documents consist of the question text (i.e. *'subject'*, *'content'*). Both types of documents were concatenated with the category fields (i.e. *'cat'*, *'maincat'*, *'subcat'*). In addition, the link graph consisting of nodes representing experts and links between questioner and answerer (provided as part of the CriES challenge) was employed as input for the HITS algorithm.

3.2 The Information Retrieval System

Support for the BM25 retrieval model [18, 19] and for the corresponding BRF approach (see Equation 1 and 2) was implemented for Lucene by one of the authors. The BM25 score for a document and a query Q is defined as:

$$\sum_{t \in Q} w^{(1)} \frac{(k_1 + 1)tf}{K + tf} \frac{(k_3 + 1)qtf}{k_3 + qtf} \tag{1}$$

---

[3] http://lucene.apache.org/
[4] http://members.unine.ch/jacques.savoy/clef/index.html

Fig. 1. Sample topic from the CriES topic set.

```
<topic lang="en">
 <identifier>3938625</identifier>
 <title>What is the origin of "foobar"?</title>
 <description>I want to know the meaning of the
   word and how to explain to my friends.</description>
 <narrative/>
 <category>Programming &amp; Design</category>
 <questioner>u1061966</questioner>
 <answerer>u25724</answerer>
</topic>
```

where Q is the query, containing terms t and $w^{(1)}$ is the RSJ (Robertson / Sparck-Jones) weight of t in Q [20]:

$$w^{(1)} = \frac{(d+0.5)/(D-d+0.5)}{(n-d+0.5)/(N-n-D+d+0.5)} \tag{2}$$

where k_1, k_3, and b are model parameters. The default parameters for the BM25 model used are $b = 0.75$, $k_1 = 1.2$, and $k_3 = 7$. N is the number of documents in the collection and D is the number of documents known or presumed to be relevant for the current topic. For the experiments described in this paper, D was set to 0, i.e. no blind relevance feedback was employed, because the number of experts and precision of our initial retrieval were presumed to be very low. n is the document frequency for the term and d is the number of relevant documents containing the term. *tf* is the frequency of the term within a document; *qtf* is the frequency of the term in the topic. $K = k_1((1-b) + b \cdot doclen/avg\_doclen)$ *doclen* and *avg_doclen* are the document length and average document length, respectively. The BM25 retrieval model has been employed for many years in evaluation campaigns such as TREC [19], but can still be considered as a state-of-the-art IR approach.

3.3 Reranking with HITS

The HITS algorithm is a link analysis algorithm for rating web pages [14]. Unlike PageRank [15], which produces a static, query-independent score, HITS produces two values for a web page: its authority and its hub value. In contrast to PageRank, HITS values are computed at query time and on results retrieved with an initial retrieval [14]. The computations are performed only on initially retrieved results, not across all linked web pages. The authority estimates the value of the content of a web page (also referred to as item in the rest of the paper, because the CriES data does not comprise web pages). In terms of expert search, the authority value indicates the quality of answers given, and indirectly the experts' level of expertise. The hub value estimates the value of its links

to other pages. Authority and hub values are defined recursively and in terms of one another. The authority value is calculated as the sum of the scaled hub values that of items linking to that item. The hub value of an item is computed by the sum of the scaled authority values of the items it links to.

To apply HITS for expert search, the expert graph is viewed as a linked graph of experts (corresponding to web pages) with directed connections (links) from questioners to answerers if the answerer provided an answer to a question.

Fig. 2. Variant of the HITS algorithm used for CriES experiments.

```
I := set of linked items
FOR EACH i IN I DO      // (initialize)
   i.auth := 1          // initial authority value of item i
   i.hub  := 1          // initial hub value of item i
FOR t=1 TO k            // run the algorithm for k steps
   a_sum := 0           // sum of authority values
   h_sum := 0           // sum of hub values
   FOR EACH i in I DO  // (update authority values)
      FOR EACH j IN i.incomingNeighbours do
                        // process items that link to i
         i.auth += j.hub
      a_sum += i.auth*i.auth
   FOR EACH i in I DO  // (update hub values)
      FOR EACH j IN i.outgoingNeighbours do
                        // process items that i links to
         i.hub  += j.auth
      h_sum += i.hub *i.hub
   FOR EACH i in I DO  // (normalize values)
      i.auth /= a_sum
      i.hub  /= h_sum
```

For the experiments described in this paper, the hub and authority values for an item are calculated with the following algorithmic steps, iterating steps (2)-(4) for k times (see also Figure 2):

(1) *Initialize:* Set the hub and authority value for each item (node) to 1.
(2) *Update authority values:* Update the authority value of each item to be equal to the sum of the hub values of each item that points to it. That is, items with a high authority value are linked to by items that are recognized as informational hubs.
(3) *Update hub values:* Update the hub value of each item to be equal to the sum of the authority values of each item that it points to. That is, items with a high hub value link to items that can be considered to be authorities on the subject.

(4) *Normalize values:* Normalize the authority and hub values by dividing each authority value by the sum of the squares of all authority values, and dividing each hub value by the sum of the squares of all hub values.

Applied to expert search, hubs can be interpreted as persons interested in a topic, and authorities can be seen as experts on a topic.

4 Experiments

The dataset consists of Q/A pairs which are maintained and verified by experts. In contrast to IR, results represent experts which may be associated with different levels of expertise; in comparison with FAQ finding, expert search focuses on looking for people most capable of providing an answer. In the simplest case, people have already provided that answer to the same or to similar questions. A graph model was provided as part of the CriES data, which consists of a directed graph representation where nodes denote topics, incoming links are questions and outgoing links represent answers.

The answer documents (A) and question documents (Q) generated from the CriES data were indexed separately. The following experimental settings were varied:

- index: retrieval on answer documents (A) or question documents (Q)
- language: no topic translation; topic translation (using Google Translate)[5]; English topics only.
- retrieval method: using standard IR (BM25); combining BM25 with HITS authority values from top 50/100 results (HITS 50/100).

The BM25 retrieval model was used with default parameters ($b = 0.75$, $k_1 = 1.2$, and $k_3 = 7$), retrieving the top 100 results for each topic. The HITS algorithm was applied on the top 50 or top 100 results retrieved by standard retrieval with the BM25 model. The HITS algorithm was run for 50 iterations ($k = 50$). The experimental setting were chosen empirically after initial retrieval experiments on CriES test data. Table 1 shows results for official and additional expert search experiments on the CriES data. The submitted runs were obtained by retrieving 100 results via IR and reranking these results with the HITS authority value. The top ten results for each topic were extracted for submission.

5 Discussion and Analysis

As described in [1], a baseline run resulting from BM25+Z-Score was generated by the organizers of the pilot challenge. This baseline experiment was based on different language-specific indexes, using Google Translate for topic translation. Z-Score normalization was employed to aggregate final results. Two different sets

---

[5] http://translate.google.com/

Table 1. Results for CriES experiments. Official experiments are set in italics.

Description	strict			lenient		
	rel_ret	MAP	P@10	rel_ret	MAP	P@10
Baseline, BM25+Z-Score	-	-	0.19	-	-	0.39
BM25, A, no topic translation	50	0.0106	0.0833	112	0.0143	0.1867
BM25, A, no topic translation, HITS (100)	56	0.0123	0.0933	241	0.0426	0.4017
BM25, A, no topic translation, HITS (50)	146	0.0584	0.2433	206	0.0393	0.3433
BM25, Q, no topic translation	38	0.0102	0.0633	109	0.0134	0.1817
BM25, Q, no topic translation, HITS (100)	46	0.0113	0.0767	251	0.0425	0.4183
BM25, Q, no topic translation, HITS (50)	128	0.0441	0.2133	183	0.0328	0.3050
BM25, A, English topics	12	0.0107	0.0800	27	0.0133	0.1800
BM25, A, English topics, HITS (50)	18	0.0139	0.1200	76	0.0569	0.5067
BM25, Q, English topics	10	0.0070	0.0667	31	0.0164	0.2067
BM25, Q, English topics, HITS (50)	9	0.0046	0.0600	74	0.0509	0.4933
BM25, A, Google Translate	41	0.0067	0.0683	71	0.0063	0.1183
BM25, A, Google Translate, HITS (50)	101	0.0409	0.1683	141	0.0266	0.2350
BM25, Q, Google Translate	23	0.0039	0.0383	61	0.0067	0.1017
BM25, Q, Google Translate, HITS (50)	75	0.0290	0.1250	119	0.0213	0.1983

of relevant judgments were provided by the organizers, corresponding to strict evaluation (experts likely able to answer are considered relevant) and lenient evaluation (experts likely able to answer and experts which may be able to answer are relevant). A comparison of our experimental results to the provided baseline reveals that our best experimental results are slightly higher in terms of P@10 (0.24 vs. 0.19 for strict, 0.42 vs. 0.39 for lenient relevance judgments). The results do not consistently outperform the BM25 baseline, and show much lower performance than the best results reported by the organizers [1]. To test if this behavior was caused by the missing topic translation, the data was analyzed in more detail with respect to the languages.

The topic set contains 60 topics. The lenient relevance assessment contains 3602 relevant entries (60.03 relevant items on average per topic), the strict assessment 1736 entries (28.93 relevant items average). Table 2 shows the distribution of languages in topics, questions, and answers. As the topics are equally distributed among four languages (15 topics per language), a more detailed analysis of the language of questions and answers was performed. This analysis shows that languages of Q/A pairs are not equally distributed among these languages, i.e. there is a bias towards English (91.3%).

Table 2. Language distribution of topics, Q/A documents and questions and answers from relevant experts.

type	language				$\sum$
	DE	EN	ES	FR	
topics	15 (25%)	15 (25%)	15 (25%)	15 (25%)	60
Q/A docs.	9219 (1.18%)	712K (91.30%)	38707 (4.96%)	19852 (2.54%)	780K
Q/A rel. strict	4054 (1.19%)	301K (88.61%)	24476 (7.19%)	10225 (3.00%)	340K
Q/A rel. lenient	9349 (1.84%)	442K (87.24%)	34410 (6.78%)	21013 (4.14%)	507K

The question index and answer index both contain 780,133 documents with the language distribution shown in Table 2.[6] While the topics are equally distributed between the four languages German (DE), English (EN), Spanish (ES), and French (FR), the majority of question and answer documents are in English.

As an additional experiment, the experimental results were calculated for the English topics only (the first 15 topics). However, there seems to be little bias towards English in relevant items compared to all items, because MAP slightly decreases from 0.1867 to 0.1800 and the number of relevant items is about a quarter of relevant items for all topics (27 vs. 112).

A comparison of retrieval on question documents and answer documents shows that results (i.e. rel_ret, MAP, and P@10) are slightly higher for IR on answer documents. A possible explanation is that answers are typically longer than questions and provide more terms to match. Thus, a lexical mismatch may be less likely. The reranking with HITS authority typically shows considerable improvement in the number of relevant and retrieved results (rel_ret), mean average precision (MAP), and P@10, more than doubling P@10 for the lenient evaluation.

There are several possible explanations of the moderate performance in comparison to the best CriES runs submitted by other participants (see, for example, the overview in [1]):

Multilinguality: All 60 topics (corresponding to queries) are equally distributed between German, English, Spanish, and French. Assuming that the questions and answers in the CriES data was similarly distributed, no topic or document translation was performed for our official experiments and all documents in their original language were organized in a single index. However, most answers in the CriES data are written in English. The baseline experiment in [1] was in contrast conducted on different language-specific indexes, combining results by Z-Score. Additional experiments on the subset of English topics and on translated topics show that there is in fact no bias towards English documents in the relevance assessments (see Table 1).

---

[6] The original number of documents reported by [1] is 780,193, but a small number of documents do not include a language code or contain invalid XML and have not been indexed.

Expert model: Experts are represented by a set of individual questions or answers. An aggregated model, i.e. combining all questions and answers into a single representation (e.g. document or weighted term vector) has not been investigated. The main reason for this is that a single category (unrelated to the current topic) may dominate in all contributions of a single user.

External resources: No additional external resources have been used for the experiments described in this paper. Standard approaches for FAQ search typically utilize resources such as WordNet [7] to bridge the lexical gap between questions and answers. However, in a multilingual problem setting, WordNet may be of limited use, because WordNet synsets contain only English terms. Multilingual resources such as EuroWordNet [21] seem to be more suitable, but suffer from a limited lexical coverage.

Link analysis: Traditional approaches for link analysis of the link graph provided in the CriES data have not been used. Instead, the HITS algorithm, which is typically applied for reranking web pages has been employed for reranking results. Interestingly, reranking initial retrieval results with HITS authority values improved performance considerably in most cases, increasing MAP and P@10 and yielding more than double the number of relevant results compared to the corresponding BM25 experiment, even for poor initial results. This result was unexpected, because the initial precision is very low for the result set retrieved with the BM25 model. Standard query expansion techniques such as blind relevance feedback also aim at improving at performance by reranking documents in a second retrieval phase, but build on the assumption that top-ranked documents in an initial retrieval phase are relevant (i.e. the initial precision is already high). If initial results have low precision, standard query expansion techniques will typically add noise instead of useful terms. Reranking results with HITS authority values improves performance despite low initial precision.

6 Conclusion and Future Work

The experiments on the CriES data show that traditional IR methods alone (i.e. the BM25 retrieval model) may not be suitable for this kind of task and social network or link analysis may be more successful. Reranking results with HITS authority values seems to improve performance even when the initial precision is low. The multilingual aspect introduced in the CriES challenge seems artificial because most contributions in the data are in English. However, experiments on the English topic subset did not show a bias.

Future work will include adding knowledge from external resources such as Wikipedia and expanding the use of categories and other metadata provided in the CriES data. Also, reranking with HITS authority scores for ad-hoc IR will be investigated.

Acknowledgments

References

1. Sorg, P., Cimiano, P., Sizov, S.: Overview of the cross-lingual expert search (CriES) pilot challenge. In: Working Notes of the CLEF 2010 Lab Sessions. (2010)
2. Surdeanu, M., Ciaramita, M., Zaragoza, H.: Learning to rank answers on large online QA collections. In: ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, The Association for Computer Linguistics (2008) 719–727
3. Harabagiu, S.M., Maiorano, S.J.: Finding answers in large collections of texts: Paragraph indexing + abductive inference. In: Proceedings of the AAAI Fall Symposium on Question Answering Systems. (1999) 63–71
4. Burke, R., Hammond, K., Kulyukin, V., Lytinen, S., Tomuro, N., Schoenberg, S.: Natural language processing in the FAQ finder system: Results and prospects. In: Proceedings of the 1997 AAAI Spring Symposium on Natural Language Processing for the World Wide Web. (1997) 17–26
5. Burke, R., Hammond, K., Kulyukin, V., Lytinen, S., Tomuro, N., Schoenberg, S.: Question answering from frequently-asked-question files: Experiences with the FAQ finder system. Technical Report TR-97-05, Dept. of Computer Science, University of Chicago (1997)
6. Wu, C.H., Yeh, J.F., Chen, M.J.: Domain-specific FAQ retrieval using independent aspects. ACM Transactions on Asian Language Processing 4(1) (2005) 1–17
7. Fellbaum, C., ed.: Wordnet. An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts (1998)
8. Dong, Z., Dong, Q.: HowNet And the Computation of Meaning. World Scientific Publishing, River Edge, NJ, USA (2006)
9. Jijkoun, V., de Rijke, M.: Retrieving answers from frequently asked questions pages on the web. In: CIKM'05, October 31-November 5 2005, Bremen, Germany. (2005) 76–83
10. Chiu, D.Y., Chen, P.S., Pan, Y.C.: Dynamic FAQ retrieval with rough set theory. IJCSNS International Journal of Computer Science and Network Security 7(8) (2007) 204–211
11. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. In: SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2006) 43–50
12. Macdonald, C., Ounis, I.: The influence of the document ranking in expert search. In: CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management, New York, NY, USA, ACM (2009) 1983–1986
13. Chang, K.H., Chen, C.Y., Lee, J.M., Ho, J.M.: EFS: Expert finding system based on Wikipedia link pattern analysis. In: IEEE explore. (2008) 631–635
14. Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 5(46) (1999) 604–632

15. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: WWW7: Proceedings of the seventh international conference on World Wide Web 7, Amsterdam, The Netherlands, The Netherlands, Elsevier Science Publishers B. V. (1998) 107–117

16. Ng, A.Y., Zheng, A.X., Jordan, M.I.: Stable algorithms for link analysis. In: SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2001) 258–266

17. Li, L., Shang, Y., Zhang, W.: Improvement of HITS-based algorithms on web documents. In: WWW '02: Proceedings of the 11th international conference on World Wide Web, New York, NY, USA, ACM (2002) 527–535

18. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3. In Harman, D.K., ed.: Overview of the Third Text Retrieval Conference (TREC-3), Gaithersburg, MD, USA, National Institute of Standards and Technology (NIST) (1995) 109–126

19. Robertson, S.E., Walker, S., Beaulieu, M.: Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track. In Harman, D.K., ed.: The Seventh Text REtrieval Conference (TREC-7), Gaithersburg, MD, USA, National Institute of Standards and Technology (NIST) (1998) 253–264

20. Robertson, S.E., Sparck-Jones, K.: Relevance weighting of search terms. Journal of the American Society for Information Science **27** (1976) 129–146

21. Vossen, P.: Introduction to EuroWordNet. In: EuroWordNet: a multilingual database with lexical semantic networks. Kluwer Academic Publishers, Norwell, MA, USA (1998) 1–17