Effective Topic Distillation with Key Resource Pre-selection

Yiqun Liu[1], Min Zhang[2], and Shaoping Ma[2]

[1] State Key Lab of Intelligent Tech. and Sys., Tsinghua University,
Beijing, 100084, China
liuyiqun03@mails.tsinghua.edu.cn

[2] State Key Lab of Intelligent Tech. and Sys., Tsinghua University,
Beijing, 100084, China
{z-m, msp}@tsinghua.edu.cn

Abstract. Topic distillation aims at finding key resources which are high-quality pages for certain topics. With analysis in non-content features of key resources, a pre-selection method is introduced in topic distillation research. A decision tree is constructed to locate key resource pages using query-independent non-content features including in-degree, document length, URL-type and two new features we found out involving site's self-link structure analysis. Although the result page set contains only about 20% pages of the whole collection, it covers more than 70% of key resources. Furthermore, information retrieval on this page set makes more than 60% improvement with respect to that on all pages. These results were achieved using TREC 2002 web track topic distillation task for training and TREC 2003 corresponding task for testing. It shows an effective way of getting better performance in topic distillation with a dataset significantly smaller in size.[1]

1 Introduction

Currently, Web Information Retrieval (IR) presents a technical challenge due to the size exploding of web document collection, which contains over 20 billion pages as of February, 2003[1]. The number of pages indexed by web search engines is increasing at a high speed, for example, Google indexed over 3.3 billion pages in September, 2003, which is about 7 times as many as what it indexed in the year of 2000[2]. How to achieve better performance with fewer pages indexed is becoming more and more interesting in web IR research. Many web search engines have adopted some techniques to identify the quality of web pages independent of a given user request, in order that they can index more high quality pages with limited resources. But these approaches such as PageRank[3] only use link structures of the web and a better estimate should require additional non-content sources of information both within a page and across different pages.

[1] Supported by the Chinese Natural Science Foundation (NO. 60223004, 60321002, 60303005)

S. H. Myaeng et al. (Eds.): AIRS 2004, LNCS 3411, pp. 129–140, 2005.

Topic distillation is a web search task to find high-quality web pages (called key resources) for a particular topic. These pages can offer users with credible information or a good entry point to several useful pages. About 78% search engine queries are related to this task according to query log analysis of Broder [4]. If key resources can be pre-selected from the whole collection, a large number of web search requests can be satisfied by indexing these pages only (because only these pages may be returned as results in topic distillation). It is possible to find several non-content features to determine whether one web page is a key resource page or not. In fact, previous TREC experiments[5][6][7] show that some non-content sources of information such as URL and link structure enhance retrieval effectiveness. Which non-content features are useful in selecting key resources and how to use them are the key components of our work.

The remaining part of the paper is constructed as follows: Section 2 gives a brief review of related works in non-content features of web pages. Section 3 compares differences between key resource pages and ordinary pages in these features. A decision tree-based key resource pre-selection algorithm is demonstrated in section 4. Section 5 describes experiment results of both key resource pre-selection and related content retrieval process. Finally come discussion and conclusion.

2 Related Work

Non-content features are sources of non-content information both within a page and across different pages, such as URLs and links. Existing studies on web page non-content features are mostly related to site finding. A site finding task is one where the user wants to find a particular site and his query names the site. It belongs to navigational search whose percentage in web search queries is over 20% according to Broder[4].

Many efforts have been made to find several non-content information sources in web site finding and several of them have proved effective. Westerveld et al. defined a web page's URL-type according to its URL length and number of dashes ('/'). They computed the probability of a page being an entry page given the type of its URL: root, subroot, path, or file. Combination of this feature and some content features helps them to obtain the best result in TREC 2001's home page finding task[8]. His colleague Kraaij et al. further proved the effectiveness of document length and in-degree in selection of home pages in 2002[9]. Craswell et al. also found that, in optimal conditions, all of the query-independent methods they studied (in-degree, URL-type, and two variants of PageRank) offered a better than random improvement on a content only baseline in site finding[10].

Key resource page has the function of providing credible information on a certain topic. It means that key resource page is different from ordinary entry page, although it is defined as entry of one key resource site according to TREC 12 web track's guideline[6]. It means that effective non-content features in site finding should be re-studied in topic distillation and new features should be introduced to separate key resource pages from ordinary pages.

3 Non-content Features of Key Resources

This section compares differences between ordinary and key resource pages in some non-content features. The features to be discussed are in-link count (in-degree), document length, URL-type, in-site out-link number and in-site out-link anchor rate; features involving self-link analysis of the site are designed specially for key resource separation. Key resource training set used here is composed of relevant qrels of TREC 2002's topic distillation task (See Section 5).

Some common-used features are discussed at first, followed by new features we proposed involving in-site out-link analysis.

3.1 Study of Common-Used Non-content Features

In-Link Count (In-Degree). Analysis of both .GOV and key resource training set shows that entry pages are different from ordinary pages in the distribution of in-degree.

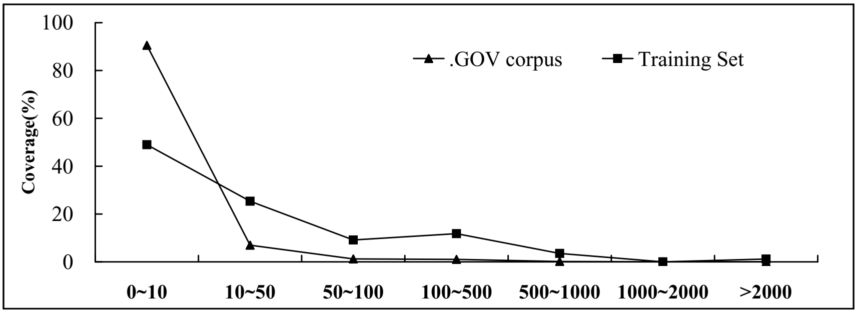

Fig. 1. In-degree distributions of key resource training set and .GOV. The category axis represents in-degree

It can be concluded from Fig. 1 that key resource pages have many more in-links than ordinary pages. The plots show that 51.03% key resource pages have more than 10 in-links; while only 9.45% ordinary pages have in-degree over 10. Key resources are high quality pages which interests many web users so a large number of web sites create hyper-links to them for convenience of these users.

URL-Type. URL-type proves to be stable and effective in site finding task according to previous TREC experiments [10][14][15]. We followed Kraaij et al. and classified URLs (after stripping off a trailing index.html, if present) into four categories: ROOT, SUBROOT, PATH and FILE. Our experiment gives the following statistics from .GOV and the training set.

There are 57.27% key resource pages in the "non-FILE" URL-type page set, which contains 11.10% pages of .GOV corpus according to Fig. 2. It means that key resource pages are likely to be "non-FILE" type. Key resource pages have

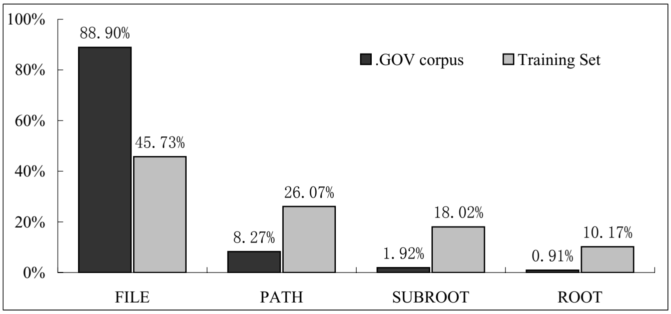

Fig. 2. URL-type distributions of key resource training set and .GOV

URLs ending with "/" or "index.html" because their authors usually use them as index pages for web sites.

The statistics also show that not all key resource pages are "non-FILE" URL-type. In fact about half of them are FILEs. NIH marijuana site entry page can be taken for example: It is a key resource for topic "marijuana", but its URL www.nida.nih.gov/drugpages/marijuana.html is obviously "FILE" type.

Document Length. Average document length (also referred to as page length) of key resource pages is 9008.02 words; it is quietly close to that of the .GOV pages (8644.62 words).

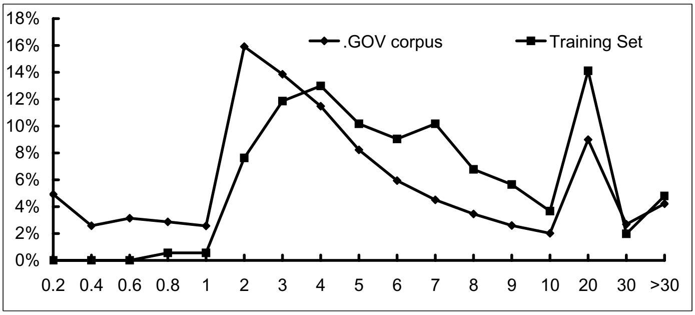

Fig. 3. Page length distributions of key resource training set and .GOV, the category axis represents "document length" in K words

The distributions in document length of .GOV and key resource pages are compared in Fig. 3. The figure shows that two page sets have similar distribution of page length, except that key resource page length would not be too short. Only 1.12% of key resource pages are shorter than 1000 words. While in ordinary pages, the percent is 16.00%. It means that pages with too few words cannot be key resource pages. This feature can be applied to reduce redundancy in the key resource page set.

3.2 New Features Involving In-Site Out-Link Analysis

In-site out-links of a certain Page A are links from A to other pages in the site where A is located. For example, the hyperlink from AIRS 2004 homepage to AIRS 2004 call for paper page is an in-site out-link of the former page. The link is located in the former page and is at the same time an in-link for the latter one.

In-Site Out-Link Number. Key resource pages have the function of linking to other informative pages of the same site. The function is like entry page's "navigation function" described in Craswell et al [11]. Key resources need more in-site out-links to finish the navigation function according to the statistics in Fig 4. Average in-site out-link number of key resource set is more than 37 compared with that of less than 18 in ordinary page set.

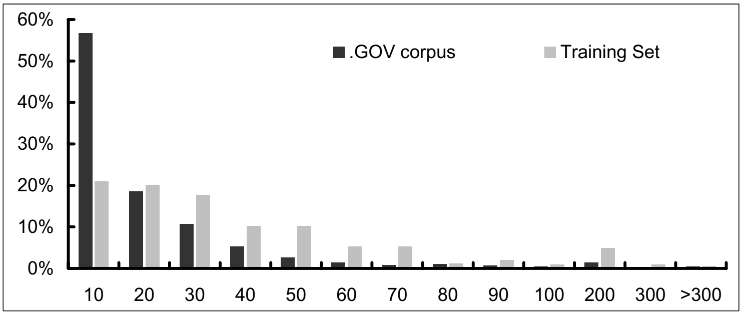

Fig. 4. In-site out-link number of key resource training set and .GOV. The category axis represents in-site out-link anchor number

A large number of in-site out-links is given to one key resource page, so it can connect directly to these informative pages in its site/sub-site. Further analysis in Section 4 shows that this attribute doesn't work as effective as in-degree and URL-type in pre-selection process, however, it helps separate key resource pages where traditional features fail.

In-Site Out-Link Anchor Rate. Key resource pages' in-site out-link anchors can be regarded as a brief review of the other pages' content in the same site. For a certain page A, in-site out-link anchors are anchors describing links from A to other pages in the same site/sub-site as A. In-site out-link anchors are located on A; that is quite different from in-link anchors which are frequently used in content analysis.

In-site out-link anchor rate is defined as:

$$rate = \frac{WordCount(in - site\ out - link\ anchor\ text)}{WordCount(fulltext)} . \qquad (1)$$

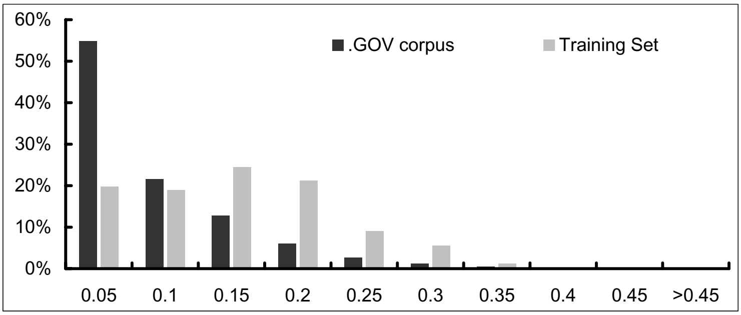

Fig. 5. In-site out-link anchor rate of key resource page training set and .GOV. The category axis represents in-site out-link anchor rate

According to Fig. 5, key resource training set has more pages with a high rate. It can be explained by the fact that key resource pages are always index pages for a site/sub-site. They are representatives of sites and in-site out-link anchors work as a summary of these sites' other pages. So one key resource page should have lots of in-site out-link anchors to introduce its site and its in-site out-link anchor rate is reasonably high.

This feature's distribution is similar with that of in-site out-link number. There are less than 24% pages in .GOV with in-site out-link anchor rate over 0.1; the percentage is about 61% in key resource training set.

4 Key Resource Page Decision Tree

Decision tree learning is adopted to combine non-content features discussed in Section 3. It is a method for approximating discrete-valued functions that is robust of noisy data and capable of learning disjunctive expressions. We choose decision tree because it is usually the most effective and efficient classifier when we have a small number of features; it also provides us with a metric to estimate feature quality in the form of information gain (ID3) or information ratio (C4.5).

ID3 algorithm proposed by Quinlan (referring to Mitchell [12]) is performed to construct the decision tree. This algorithm uses information gain to decide which attribute should be selected as the decision one for the root node of the decision tree. Information gain is a statistical property that measures how well a given attribute separates the training examples according to their target classification.

For this particular problem of selecting key resource pages, non-content features with continuous values should be discretized in advance. Attributes range should be partitioned into two or several intervals using a single or a set of cut points (thresholds). One fixed threshold is chosen in our method and after discretization process each non-content feature has boolean values.

According to ID3 algorithm described in [12], in-degree with the most information gain should be chosen at the root node (cf. Table 1). Then the process of selecting a new attribute and partitioning the training examples is repeated

Table 1. Information gain with different attributes

Attribute	Information gain
In-degree	0.2101
URL-type	0.1981
Document length	0.0431
In-site out-link number	0.0796
In-site out-link anchor rate	0.0988

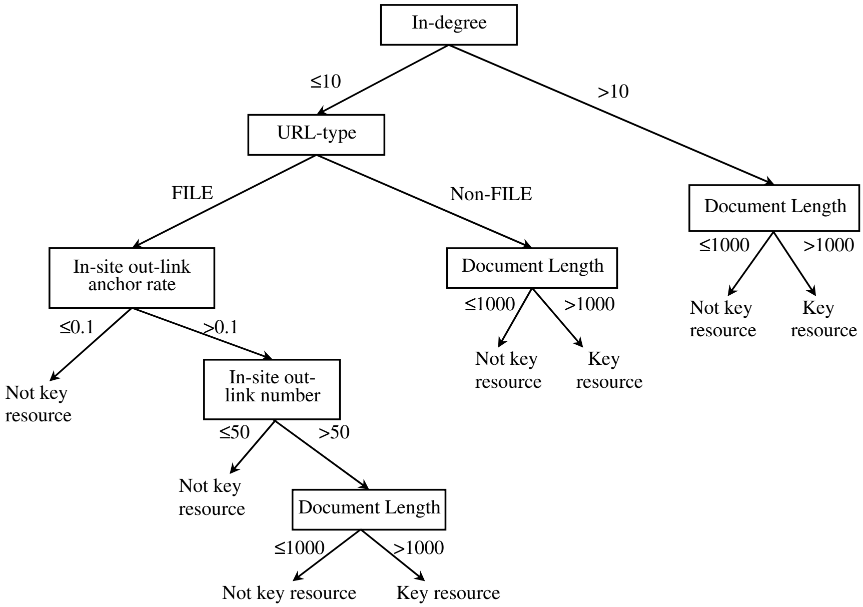

Fig. 6. Key resource decision tree constructed with ID3

for each non-terminal descendant node, with only the examples associated with the node.

In the construction process, non-content attributes are ranked by their information gains while choosing the root node, as shown in Table 1.

In-degree and URL-type are good classifiers to separate key resources from ordinary pages and their information gains are almost the same. Information gain of document length is quite low, but according to Section 3.3 it can be used to keep out redundancy.

The two features obtained from in-site out link analysis are important in decision tree learning. Sub tree with these two features is used to select key resources from the pages whose URL type are FILE and in-degree are less than 10. The percentage of such kind of pages in .GOV is 68.53%, and there are 29.38% of key resource pages among them. Without in-site out-link analysis, it is impossible to get these key resource pages separated.

When one example set is mainly composed of pages with the same target attribute value or all attributes have been tested, the construction process ends. Finally the decision tree shown in Fig. 6 is constructed. With this decision tree, any page in .GOV page set can be judged whether it belongs to the key resource set or not.

5 Experiments and Discussions

There are two methods to evaluate the effectiveness of key resource pre-selection. First is the direct evaluation: if the result set selected by the decision tree covers a large number of key resources with a small set size, pre-selection can be regarded as a success. Second, topic distillation on the result set should get high performance although the set size is smaller than the whole page set: It is called indirect way of evaluation. Both direct and indirect evaluations are involved in our experiments.

5.1 Training Set and Test Set

The key resource page training set is based on relevant qrels given by TREC 2002's topic distillation task. The task is to find key resource pages for certain topics but there may be several high-quality pages from one single web site. According to a better key resource definition in TREC 2003, we use entry page to represent all pages in this kind of high-quality site/sub-site.

For example, the entry page of ORI scientific misconduct sub-site whose URL is http://ori.dhhs.gov/html/misconduct/casesummaries.asp is used to replace all qrels from the same sub-site for the topic "scientific research misconduct cases" (Topic No. 559). Because it is the entry page of the site and it has in-site out-links to the following relevant qrels:

http://ori.dhhs.gov/html/misconduct/elster.asp
http://ori.dhhs.gov/html/misconduct/french.asp
http://ori.dhhs.gov/html/misconduct/hartzer.asp
(another 21 qrels from the same site are omitted here)

TREC 2003's topic distillation task is to find as many key resource site entry pages as possible. The task's 50 queries are from log analysis of web search engines and NIST gets credible qrels with pooling technology. The topics and relevant qrels are directly used as the test set.

5.2 Key Resource Coverage of the Result Set

When fixed threshold values vary in the discretization process described in Section 4, result sets with different size and key resource coverage are built correspondingly. Statistics of several result sets are shown in Fig. 7, from which two important conclusions can be drawn:

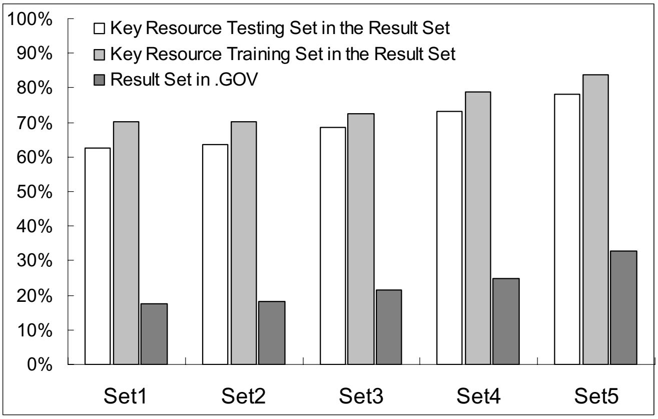

Fig. 7. Key Resource coverage and the result set size

1. With the decision tree constructed in Section 4, it is possible to cover about 70% test set pages (key resource pages) with about 20% of .GOV size. The result set can cover a high percentage of key resources with a small amount of pages because non-content features which we adopt are fit for the job of selecting key resources. There are still about 30% key resource testing set pages which the result set doesn't include. It means the up-limit of this key resource pre-selection method is about 70%. However, topic distillation performance is less than 13% with the measure of precision at 10 currently according to TREC 11 and TREC 12 reports[5][6]. The up-limit doesn't depress the performance too much.
2. Key resource coverage increases with the key resource set amount. It is obvious that 100% coverage is got with the result set equaling .GOV, so there must be a trade off between the result set size and key resource testing set coverage.

5.3 Topic Distillation on the Key Resource Set

Experiments in this Section are based on a key resource set with 24.89% pages in .GOV and 73.12% key resources (Set 4 in Fig. 7). BM2500 weighting and default parameter tuning described in[14] are performed in all experiments. Topics and relevant qrels in TREC 2003 topic distillation task are used for testing. In-link anchor retrieval is evaluated together with full text retrieval both on .GOV corpus and on the key resource set, because in-link anchor proves to be an effective source for topic distillation in [6]. These results are also compared with TREC 2003 best run to validate effectiveness of our method in Fig. 8.

In-link anchor retrieval performs much better than full text retrieval on .GOV (46% improvement in R-precision and 42% in P@10). It accords with the previous conclusion by Craswell et al[11] that BM25 ranking applied to link anchor documents significantly outperforms the same ranking method applied to document content. However, even anchor text retrieval on .GOV gets much worse

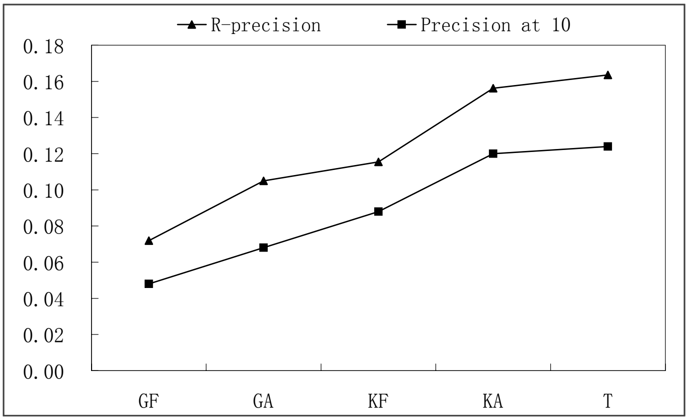

Fig. 8. Text retrieval on .GOV and Key resource set compared with TREC 2003 best run, *GF*: Full text of .GOV corpus, *GA*: In-link anchor text of .GOV corpus, *KF*: Full text of Key resource page set, *GA*: In-link anchor text of Key resource page set, *T*: Best run in TREC 2003 web track

performance than TREC 2003 best run. It can be explained by the fact that these results don't make use of any kind of non-content information.

Fig. 8 also proves that retrieval on key resource site get much better performance than that on .GOV. Similar with retrieval results on .GOV, anchor text retrieval on key resource set outperforms full text retrieval by about 35% in both evaluation metrics. However, from Fig. 8, we can see both full text and anchor text retrieval gain much progress on key resource set. Anchor text retrieval on key resource set achieves almost the same ranking as TREC 2003 best run (cf. Table 2).

These experiments are not biased towards the key resource pre-selection method. TREC 2003 topic distillation topics and relevant qrels are not used for training in any way. BM2500 ranking based on link anchor of the key resource set gets the 2nd highest R-precision and 4th highest Precision at 10 according to TREC 2003 topic distillation task results. Further improvement is expected if better stemming and parameter tuning technology are introduced.

Table 2. Anchor text retrieval on different data set compared with TREC 2003 topic distillation best results

Evaluation metric	Precision at 10	R-Precision
Whole page set	0.0680	0.1050
Key resource page set	0.1200	0.1562
Best run in TREC 2003 (highest P@10)[6]	0.1280	0.1485
Best run in TREC 2003 (highest R-precision)[6]	0.1240	0.1636

The key resource set works well with only 20% amount of the whole page set. It proves that key resource set has included a large number of high quality pages. This set can be used for topic distillation instead of the whole page set.

6 Conclusions and Further Work

In this paper, an effective topic distillation method with key resource pre-selection is proposed. Experiments prove that information retrieval on this pre-selected key resource page set gets much better performance than that on the whole collection. We can conclude that:

1. A small sub-set of web pages which contains most key resource pages can be built with key resource pre-selection technology. This sub-set contains only about 20% of the whole collection, but covers more than 70% key resources.
2. Decision tree learning is fit for the job of selecting key resources using non-content features. ID3 also provides us with a credible metric of non-content feature effectiveness.
3. Pre-selection of key resource pages works well for topic distillation according to experiment results. Full text retrieval on key resource page set gets more than 60% improvement comparing with that on the whole collection. Anchor text retrieval works as well as TREC 2003's best run on this sub set.

This research may help web search engines to index fewer pages without losing performance in topic distillation. It can also be used to evaluate web page quality query-independently based on whether it is a key resource or not.

However, a great many aspects of the key resource set are to be investigated in future work: Will we get better results if existing link analysis methods such as HITS and PageRank are performed on this set? In-site out-link analysis has proved effective in key resource selection, how well does retrieval on in-site out-link anchor instead of in-link anchor in the key resource set? Is it possible to find the best trade off point between key resource set size and key resource coverage?

References

1. Lyman, Peter and Hal R. Varian, "How Much Information", 2003. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on April 2th, 2004
2. Danny Sullivan, Search Engine Sizes. In search engine watch website; September 2, 2003. Online at: http://searchenginewatch.com/reports/article.php/2156481
3. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In Computer Networks and ISDN Systems (1998) 30(1–7): 107–117
4. Andrei Broder: A taxonomy of web search. SIGIR Forum(2002), Volume 36(2):3–10
5. N. Craswell, D. Hawking: Overview of the TREC-2002 web track. In NIST Special Publication 500-251: The Eleventh Text REtrieval Conference (TREC 2002)
6. N. Craswell, D. Hawking: Overview of the TREC-2003 web track. In NIST Special Publication 500-255: The twelfth Text REtrieval Conference (TREC 2003):78–92

7. E. M. Voorhees and D. K. Harman, editors. The Tenth Text Retrieval Conference (TREC-2001), volume 10. National Institute of Standards and Technology, NIST, 2001
8. T. Westerveld, D. Hiemstra, W. Kraaij. Retrieving Web Pages Using Content, Links, URLs and Anchors. In NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001) (2001) 663–672
9. W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In 25th ACM-SIGIR conference on research and development in information retrieval (2002) 27–34
10. Nick Craswell and David Hawking. Query-independent evidence in home page finding. In ACM Transactions on Information Systems (2003) 21(3): 286–313
11. Nick Craswell, David Hawking and Stephen Robertson. Effective Site Finding using Link Anchor Information. In 24th ACM-SIGIR Conference on Research and Development in Information Retrieval. (1998) 250–257
12. Tom M. Mitchell. Chapter 3: Decision Tree Learning, in Machine Learning (ISBN 0-07-115467-1). Pages 55-64, McGraw-HILL INTERNATIONAL EDITIONS, 1997
13. Van Rijbergen.: Information Retireval. Butterworths, London, 1979
14. S. E. Robertson, S. Walker, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In NIST Special Publication 500-225: Overview of the Third Text Retrieval Conference (1994) 109–127