Managing Quality of Information Retrieval for Effective Knowledge Management

Naresh Kumar Agarwal, Danny C. C. Poo and Jie Mein Goh

School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543
{naresh, dpoo, gohjm}@comp.nus.edu.sg

Abstract. There have been continual efforts to improve user-experience and information-retrieval through quality searching techniques. Search methods employed by search engines can be broadly seen to suffer from two patterns: a) relevant information is present but comes with much irrelevant information b) relevant information is present but not in the required quantity; relevant information exists in the Internet not indexed by the search engine in use. We examine an approach to improve search quality for different categories of users through 1) information localization and 2) specialty-search. Ease in browsing is facilitated by 3) taxonomy-based presentation/classification and 4) multiple views of the taxonomy. Relevant search is also aided by 5) the usage of cues. Our case study is based on an online portal for the Singapore Education Community. An outcome of the project is to demonstrate how improved navigation and search quality can enhance the efficacy of knowledge management in an organization.

Keywords: Knowledge Management, Information Retrieval, Search Engine, Search Quality, Localization, Specialty Search, Education, Taxonomy, Cues.

1 Introduction

1.1 The Knowledge Management Problem – Retrieving Relevant Information

A Collection of data is not information, a collection of information is not knowledge, a collection of knowledge is not wisdom and a collection of wisdom is not truth (observation by Fleming, Neil D. 1996). Knowledge Management plays an important role in an organization by facilitating the capture, storage, transformation and dissemination of information. Organizations have been experimenting with knowledge management to try and promote efficiency, improve profits, be innovative and have a competitive advantage and sometimes, simply to survive (Wigg 1997; Prusak 1997; Hendriks and Virens 1999; Loucopoulos and Kavakli 1999; Davenport and Prusak 2000; Gao et al. 2002). However, Knowledge Management remains a challenge, with a typical mid-sized organization today accumulating more information in a week than the whole of mankind did between the years 1 and 1500 A.D. Masses of documents, e-mails, databases, images, audio and video recordings form vast repositories of information assets to be tapped by employees, partners, customers and other stakeholders (Papadopoullos, Alkis 2004). Trying to make sense of all this data and being able to effortlessly retrieve relevant information forms a core part of Knowledge Management.

1.2 Inadequacies of Searching Techniques in meeting Search Quality

There have been continual efforts to improve user-experience and information-retrieval through quality search and search-optimization techniques (Shapiro and Lehoczky 2003; Search Engine Watch 2005; Notess 2004; Search Engine Guide 2005), and a lot of people today rely on the Internet to search for information related to their work.

However, a particular search engine may not be effective in meeting all the needs of a particular searcher, affecting search quality as a result. Lawrence and Giles (1998) conducted a study of 6 World Wide Web search engines and concluded that the coverage of any one engine is significantly limited: No single engine indexes more than one-third of the "indexable Web". Notess (2003) ran a search-engine analysis and describes how search engines differ in terms of relative size of search-engine databases, age/freshness of databases, reflection of web-growth, database overlap, unique hits, dead links, etc. Abilock (2005) outlines how one is better off switching from one search engine to another depending on need (to get few good hits fast, to compare results quickly, to search on misspelt words, opinions, primary sources, etc.). Johnson (2003) describes how a leading search engine, though good, leads to undesired results at times due to the backfiring of optimized algorithms it relies on for information requests.

2 Information Retrieval Relevance Model

Search methods employed by search engines can be broadly seen to suffer from one or both of two patterns: a) where relevant information is present but comes with much irrelevant information, affecting the quality of retrieved data b) where relevant information is present but not in the required quantity. There is still sufficient relevant information in the World Wide Web not indexed by the particular search engine that the searcher is using.

In Figure 1 below, we propose a model depicting the amount of relevant information returned (along with non-relevant information) as a result of a search query in the Internet using a particular search engine. The outer dotted circle *A* represents the total amount of information in the Internet which can be indexed by search engines. The lined circle *B* represents the amount of information actually indexed by a particular search engine. The area *C+E* represents the amount of information relevant to the search query present in the Internet. The area *C+D* represents the result of a search query returned by a particular search engine, from which only the area represented by the black circle C is relevant to the searcher.

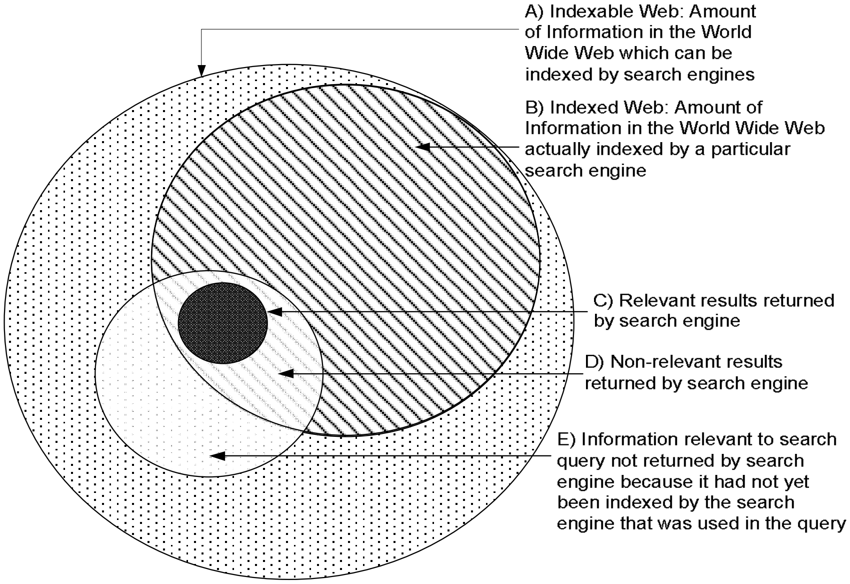

A) Indexable Web: Amount of Information in the World Wide Web which can be indexed by search engines
B) Indexed Web: Amount of Information in the World Wide Web actually indexed by a particular search engine
C) Relevant results returned by search engine
D) Non-relevant results returned by search engine
E) Information relevant to search query not returned by search engine because it had not yet been indexed by the search engine that was used in the query

Fig. 1. Information Retrieval Relevance Model

2.1 Definition of Search Quality

Information Retrieval performance is generally assessed using standard quantifiable metrics of precision[1] (C/(C+D) in Fig. 1), recall[2] (C/(C+E) in Fig. 1) and F-Measure[3] (van Rijsbergen 1979). From the query results, the user typically reads only the first k documents and not all the documents (Davis 2002). Komarjaya, Poo and Kan (2004) use precision of the top k documents or precision-at-k as their performance metric. Also, users are usually more interested in the precision of the results displayed in the first page of the list of retrieved documents (Kobayashi and Takeda 2000). We propose a definition for search quality as follows:

> ***Search quality*** *is a measure of relevant links returned in the first k links of search results.*
> *A* ***high quality search result*** *would be one that would totally satisfy the searcher, with results closely matching what the searcher hoped/expected to find while initiating the search.*

[1] proportion of retrieved material that is actually relevant

[2] proportion of relevant material actually retrieved in answer to a search request

[3] combines recall (r) and precision (p) with an equal weight in the form F-measure = 2rp / r+p

The *k* relevant links (typically first two pages of search results) could lie anywhere within the region C+E in Fig. 1.

3 Our Approach to improve Search Quality

Access to vast stores of information can no longer be ensured by simply launching the browser, going to one's search page and typing a word or two in the hopes of locating one of the oft-required needles from the haystack. Search needs far exceed the simple requirement to locate a document with a given word in it. Search and classification results must satisfy four basic categories of users – a) who have a good idea what they are looking for, know that a given document or piece of data exists, and simply need to locate it b) who need information about a topic they are knowledgeable about and are therefore in data-gathering mode c) who need information about a topic they are NOT familiar with in preparation for starting a new project d) who need a very specific answer to a specific question (Papadopoullos 2004). As Papadopoullos states, there is a need to move beyond "finding the needle in a haystack" to "connecting the dots" among various pieces of information.

Instead of adopting a 'one size fits all' information retrieval model, we propose to solve some of the problems highlighted in Fig. 1 and help 'connect the dots' through a number of ways 1) localizing search to a limited geographical context 2) specialty search engines 3) taxonomy-based presentation and classification 4) providing multiple views of the taxonomy and 5) usage of cues. Using a combination of these methods would lead to better search quality and greater satisfaction for the different categories of users highlighted above.

3.1 Information Localization

From a searcher's perspective, one way of improving the relevance of search results is through geographical localization (In Fig. 1, decreasing B to encapsulate more of C+E). While global information is useful and widely available on the Internet, there is a still a long way to go before information available from the web can be useful to a searcher from a local perspective. In the first of his five-part series, Sullivan (2003) explains how local commerce searching can often be a disappointing experience on the web's major search engines (a role performed well by the low-tech local Yellow Pages). There are signs that the major search engines are joining the push for local search (one of the motivators being the lure to tap on to the revenue generated through local advertising) and are allowing searches for local content and business data like maps, driving directions, weather, people search and movies (Sullivan 2003; Sterling 2004; Mara 2004; Notess 2005). While most search engines and major portals are pushing to provide country-specific sites, there is still some way to go before locally-relevant data ranging from education, governance, lifestyle, etc can be easily available and also satisfy the searcher's quest for quality search.

3.2 Specialty Search Engines

While localization refers to providing information more relevant to a geographical region (e.g. Singapore or Eastern Europe or Madrid), specialty search engines help provide information specific to an area or domain e.g. a search engine to be used exclusively by doctors or the medical community, saving them from having to weed out basic health/fitness information meant for the lay man and helping them focus on specific issues like the latest advances in medical science or medical job opportunities. If the big search engines are unable to deliver comprehensive access to the entire web, perhaps the time has come for more focused sites to offer near-comprehensiveness in their own chosen fields (Kawin 2003) i.e. decrease B to encapsulate more of C+E (Fig. 1). While general search engines cover the breadth of information on the Internet, specialized search engines provide access to the so-called 'deep' or 'invisible' (Khoussainov and Kushmerick 2003; Battelle 2004) and help you find more than just web pages or websites. They are also called topical search engines, 'vertical' search engines or 'vortals' and help you search through specific types of listings in different areas (Sullivan 2000). With tools to help people build their own specialized search engines (Vortaloptics 2004, etc.), vertical portals can become powerful vehicles for improved search quality and user satisfaction.

3.3 Taxonomy-based Presentation and Classification

To help different categories of users make sense of different pieces of information and connect the dots, traditional search modes (major search engines) must be complemented by information localization (narrowing the results), specialty search as well as navigable search results (taxonomies) that help users browse rather than search. Taxonomies provide a subject-based classification that arranges the terms in a controlled vocabulary into a hierarchy. This allows related terms to be categorized into meaningful frameworks that add some logical structure that humans can rapidly navigate to find high concentrations of topic-specific, related information (Papadopoullos 2004; Garshol 2004) and helps increase area C in Fig. 1.

3.4 Multiple-Views of the Taxonomy

Taxonomies are flexible structures, and can be developed to cover many different topics to any desired level of granularity. Even more powerful are dynamic classifications that allow search results to be organized in real-time into classification views that are selected by the user in order to view information from various perspectives (Papadopoullos 2004). In Fig. 1, this will help the user quickly zero-in on the results from area C. Similarly, faceted classifications work by identifying a number of facets[4] into which the terms are divided e.g. classifying by color,

[4] Facets can be thought of as different axes along which documents can be classified (faceted classification), and each facet contains a number of terms. This would then describe the document from many different perspectives (Garshol 2004). Taxonomy is a type of facet in which the headings are arranged into a hierarchy (FacetMap 2003).

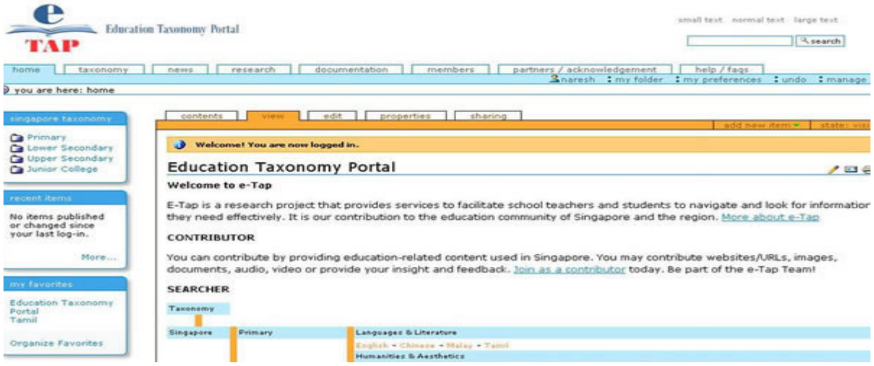
geography, subject, etc. FacetMap 2003 uses a demo to classify wines into different facets - variety, region and price.

3.5 Usage of Cues

The notion of context has been introduced to enhance search tools and refers to a diverse range of ideas from specialty-search engines to personalization. It has been found that incorporating such contextual cues from static content sources, dynamic content sources, static collaborative sources and dynamic collaborative sources can help to increase the relevance of information (Goh and Poo 2004) and enhance search quality by increasing the set of relevant results (area C in Fig. 1) and decreasing non-relevant results (area D in Fig. 1).

4 Case-Study: Education Taxonomy Portal (ETaP)

We have developed an online portal "*Education Taxonomy Portal (ETaP)"* (accessible from http://etap.comp.nus.edu.sg – see Fig. 2) as a case study to demonstrate the effectiveness of the five attributes discussed above in enhancing search quality. ETaP is a portal for the Singapore Education Community and provides services to facilitate school teachers and students to contribute, search, navigate and retrieve education-related content effectively. Information retrieved is specific to users' local needs while enabling them to contribute and share their contents. A taxonomy based on the prescribed education curriculum helps in easy browsing.



No items published or changed since your last log-in.

Welcome! You are now logged in.

Education Taxonomy Portal

Welcome to e-Tap

E-Tap is a research project that provides services to facilitate school teachers and students to navigate and look for informatior they need effectively. It is our contribution to the education community of Singapore and the region. More about e-Tap

CONTRIBUTOR

You can contribute by providing education-related content used in Singapore. You may contribute websites/URLs, images, documents, audio, video or provide your insight and feedback. Join as a contributor today. Be part of the e-Tap Team!

SEARCHER

Fig. 2. Snapshot of ETaP – Education Taxonomy Portal (http://etap.comp.nus.edu.sg)

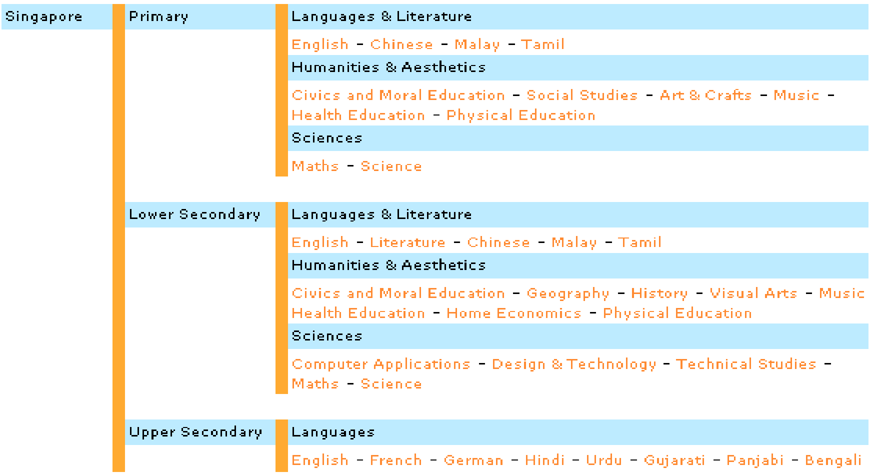
Teachers looking in the Internet for teaching materials and information relevant to their courses are almost always presented with a huge amount of data. Gathering required information is a long-drawn and time-consuming process running into hours. Students who want to search for information for project work or to supplement their course materials are similarly presented with a huge array of non-relevant data.

Parents, education policymakers, tutors, coaching/tutoring agencies, all go through the same fate.

There are many education-related professionals, teachers and schools who, in the past couple of years, have compiled their own frequently-used education material as well as useful links gathered while browsing. Different organizations/individuals have their own small repositories. The project aims to provide a country-wide repository for gathering such material (websites, images, audio, video, journals, etc) and classifying it in different categories for quality search.

In ETaP, we facilitate localization and search quality by narrowing the education-related content to the local Singapore context. Retrieved education content is current and conforms to the syllabus prescribed by the Singapore education council. As the local teachers and students are contributors as well as searchers, the relevance of content for the local education sector is significantly enhanced. Subsequently, the portal will be expanded to include countries and schools in the Asean region, while keeping data relevant to the chosen countries.

ETaP provides specialization by focusing on the Education domain. The portal will eventually be expanded to gather relevant education-related material from major search engines and combine the result set based on user needs (specialty search).



Singapore

Primary

Languages & Literature

English - Chinese - Malay - Tamil

Humanities & Aesthetics

Civics and Moral Education - Social Studies - Art & Crafts - Music - Health Education - Physical Education

Sciences

Maths - Science

Lower Secondary

Languages & Literature

English - Literature - Chinese - Malay - Tamil

Humanities & Aesthetics

Civics and Moral Education - Geography - History - Visual Arts - Music Health Education - Home Economics - Physical Education

Sciences

Computer Applications - Design & Technology - Technical Studies - Maths - Science

Upper Secondary

Languages

English - French - German - Hindi - Urdu - Gujarati - Panjabi - Bengali

Fig. 3. View of section of ETaP Taxonomy

Ease of browsing is facilitated by taxonomy-based classification and presentation, and information is organized based on the needs, perspective and vocabulary of primary and secondary schools and junior colleges in Singapore (see Fig. 3). The scope will later be expanded to include other regional countries.

Multiple views of the taxonomy or faceted classification can be applied in various ways – student’s view, teacher’s view, syllabus view (Fig. 3), classified according to a knowledge area, etc. Each view can be further classified based on date/freshness, subjects registered by student, etc. Multiple tags could be associated with a single object that would enable it to be classified under different taxonomy views that are created on the fly, depending on the navigation pattern of the user during a particular browsing/searching session.

To further improve search quality, we intend to apply the different types of contextual cues (see Section 3.5) in the system. Static content sources can be added by utilizing a database from participating schools containing users’ information such as their names and majors. Dynamic content sources can be captured using a system that logs the users’ actions. Users of the system can create a record of users whom they know so as to utilize the contextual cues that can be obtained from static collaborative sources. With such information, dynamic collaborative sources can also be obtained by matching the actions of the users with those of users with similar interests. The contribution of each type of contextual cues will help to improve the search quality of the system.

The portal aims to help teachers, students, parents and all associated with the education community in Singapore perform quality search and be better satisfied with their search results. Improved navigation and search quality should give rise to more innovation and effectiveness, and enhance the efficacy of Knowledge Management in Singapore. The scope will subsequently be expanded to other countries and regions. Experiments will be conducted to measure the search quality, which will be reported in due course. ETaP is available free for everyone’s use.

5 Conclusion

Information Retrieval forms a core part of knowledge management. There is a need to move beyond 'finding the needle in a haystack' to 'connecting the dots' among various pieces of information. In this quest to connect the dots, a 'one size fits all' model for information retrieval is inadequate.

We proposed an information retrieval relevance model showing that search methods typically suffer from one of both of two patterns 1) relevant information comes packaged with non-relevant information 2) user does not get all relevant information present as it may not have been indexed by the search engine in use. We defined search quality as a measure of relevant links returned (in relation to what the searcher had in mind before initiating the search) in the first k links of search results and the level of satisfaction the searcher feels with the search results.

Through our case study, we also hope to show that using a combination of information localization, specialty search, taxonomy-based classification and presentation, multiple, dynamic views of the taxonomy and the usage of cues is the need of the hour to improve search quality and aid in effective information retrieval for knowledge management.

References

1. Abilock, D. (2005) “Information Literacy: Search Strategies - Choose the Best Search for your Information Need”, *NoodleTools*, [http://www.noodletools.com/debbie/literacies/information/5locate/adviceengine.html].
2. Battelle, J. (2004) “GlobalSpec: Domain Specific Search and the Semantic Web”, *John Battelle's Searchblog* [http://battellemedia.com/archives/000519.php]
3. Davenport, T.H. and Prusak, L. (2000) “Working Knowledge: How organizations Manage What They Know”, *Harvard Business School Press*, Boston.
4. Davis, E. (2002) “Web search engines: Retrieval” [http://www.cs.nyu.edu/courses/fall02/G22.3033-008/lec5.html]
5. FacetMap (2003) “FacetMap: Your Home for Faceted Classification Tools” [http://facetmap.com]
6. Fleming, N.D. (1996) “Coping with a Revolution: Will the Internet Change Learning?”, *Occasional Paper for Faculty*, Lincoln University, Canterbury, New Zealand [http://www.vark-learn.com/documents/Information_and_Knowle.pdf].
7. Gao, F., Li, M. and Nakamori, Y. (2002) “Systems thinking on knowledge and its management: systems methodology for knowledge management”, *Journal of Knowledge Management*, Vol.6, No.1, pp.7-17.
8. Garshol, L.M. (2004) “Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all”, *Ontopia* [http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html#N828]
9. Goh, J.M., Poo, D.C.C. and Chang, K.T.T (2004) “Incorporating Contextual Cues into Electronic Repositories”, *Proceedings of the Eighth Pacific-Asia Conference on Information Systems*, Shanghai, China, 8-11 July 2004, pp. 472-484.
10. Hendriks, P. and Virens, D. (1999) “Knowledge-based systems and knowledge management: Friends or Foes?”, *Information & Management*, vol.35, pp.113-125.
11. Johnson, S. (2003) “Digging for Googleholes - Google may be our new god, but it's not omnipotent”, *Slate*, Washingtonpost.Newsweek Interactive Co. LLC [http://slate.msn.com/id/2085668/]
12. Kawin (2003), “Why vortal”, Vortalbuilding.com [http://www.vortalbuilding.com/whyvortal.html]
13. Khoussainov, R. and Kushmerick, N. (2003) “Learning to Compete in Heterogenous Web Search Environments”, *Proceedings of the Eighteenth International Joint Conference on Artifical Intelligence* (IJCAI-03)
14. Kobayashi, M. and Takeda, K. (2000) “Information retrieval on the web”, *ACM Computing Surveys* 32, pp. 144-173
15. Komarjaya, J., Poo, D.C.C. and Kan, M.Y. (2004) “Corpus-Based Query Expansion in Online Public Access Catalogs”, *ECDL 2004*, LNCS 3232, pp. 221-231

16.Lawrence, S. and Giles, C.L. (1998) "Searching the World Wide Web", *Science*, Vol 280, 3 April 1998 [www.sciencemag.org]

17.Loucopoulos, P. and Kavakli, V. (1999) "Enterprise Knowledge Management and Conceptual Modelling", *Lectures Notes in Computer Science*, Vol.1565, pp.123-143.

18.Mara, J. (2004), "Local Search: Working Hard for the Money", ClickZ News [http://www.clickz.com/news/article.php/3333601]

19.Notess, G.R. (2003) "Search Engine Statistics", *Search Engine Showdown* [http://searchengineshowdown.com/stats/]

20.Notess, G.R. (2004) "Search Engine News", *Search Engine Showdown* [http://searchengineshowdown.com/]

21.Notess, G.R. (2005) "Internet Search Engine Update", *Online Magazine*, Vol. 28 No. 6, Nov/Dec 2004, Information Today, Inc. [http://www.infotoday.com/online/nov04/SearchEngineUpdate.shtml]

22.Papadopoullos, A. (2004) "Answering the Right Questions about Search", EContent Leadership Series - Strategies for...Search, Taxonomy & Classification, *Supplement to July/August 2004 EContent and Information Today*, pp. S6-S7 [http://www.kmworld.com/publications/whitepapers/crm/EC_Search_04.pdf].

23.Prusak, L. (1997) "Knowledge in Organisations", Butterworth-Heinemann, Oxford.

24.Search Engine Guide (2005) "Current Search Engine News", K Clough, Inc [http://www.searchengineguide.com/searchenginenews.html]

25.Search Engine Watch (2005), Jupiter Media Corporation [http://searchenginewatch.com/].

26.Shapiro, Y. and Lehoczky, E. (2003) "Search Engine Optimization", SearchEngines.com [http://www.searchengines.com/intro_optimize.html].

27.Sterling, G. (2004) "Local Search: The Hybrid Future", *Search Engine Watch* [http://searchenginewatch.com/searchday/article.php/3296721]

28.Sullivan, D. (2000) "The Vortals are coming! The Vortals are coming!", *SearchEngineWatch* [http://searchenginewatch.com/sereport/article.php/2162541]

29.Sullivan, D. (2003) "Local Search Series", Parts 1-5, *SearchEngineWatch* [http://searchenginewatch.com/searchday/article.php/3111631]

30.Wigg, K.M. (1997) "Knowledge Management: Where Did It Come From and Where Will It Go?", *Expert System With Application*, Vol. 13, No. 1, pp. 1-14.

31.van Rijsbergen, C.J. (1979) "Information Retrieval" (second edition), in London: Butterworths, Chapter 7 Evaluation [http://www.dcs.gla.ac.uk/~iain/keith/]

32.Vortaloptics (2004), *Vortaloptics* [http://www.vortaloptics.com]