MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Contextual Annotation of Web Pages for Interactive Browsing

Erik van Mulligen[a,b], Mario Diwersy[c], Bob Schijvenaars[a,b], Marc Weeber[a], Christiaan van der Eijk[a], Rob Jelier[a], Martijn Schuemie[a], Jan Kors[a], Barend Mons[a]

[a]Dept of Medical Informatics, Erasmus University Medical Center Rotterdam, The Netherlands
[b]Collexis BV, Geldermalsen, The Netherlands
[c]Syynx WebSolutions GmbH, Frankfurt, Germany

Abstract

With the information on the World Wide Web and in specialized databases exploding, researchers and physicians are in dire need to browse efficiently though the large corpus of information resources in their field of interest. The focus is not any longer to find everything related to your interest, but it shifts to zooming in, based on context and expanding again in neighboring knowledge domains. This paper describes an attempt to develop a completely new, interactive way of browsing distributed corpora of information without the need for multiple different queries in different information resources.
Classical search engines generally treat search requests in isolation. The results for a given query are identical, and do not automatically take on board the context in which the user made the request. The system described here explores implicit contexts as obtained from the document that the user is reading. The new approach merges the searching and browsing into one combined "read-and-search" mode and alleviates the shift users are normally forced to between searching and reading.

Keywords:

Information Retrieval System, Hypertext, User-Computer Interface

Introduction

In a classical search, contextual information to narrow or widen a search may be provided by the user in the form of keywords added to the query. For example, a user looking for the treatment options for a particular disease may add keywords such as "drug" or "treatment". However, this approach frequently has a very deleterious side effect: most search engines are based on Boolean algebra. Boolean queries can be effectively used to retrieve those documents that fulfill the query from a large set of documents. However, formulation of a multiple keyword query is difficult and assumes that the user is familiar with both the exact syntax and the right terms, including synonyms and homonyms. Query overspecification – adding spurious search terms – leads to low recall and query underspecification – too few search terms – to low precision, and to "the million hits syndrome". Several alternatives have been developed to overcome these Boolean query problems. These all map into the categories of rule based extension of the Boolean algebra to include fuzzyness and new approaches such as approximative matching based on the vector space model [1].

Obtaining context

Several existing projects have attempted to enrich queries with contextual information. The Inquirus 2 project [2] at the NEC research institute asks the user to explicitly indicate the category – "personal homepage", "research paper", or "general introductory information" – of the information searched for. This categorization is used as context for the search. The Watson project [3] seeks to automatically infer the context. In this project, a document is analyzed with a heuristic term weighting algorithm, aiming to find terms that are representative of the content of the documents. These terms are used to modify the query entered by the user. Autonomy's Kenjin program [4] automatically suggests content from the web, based on the documents a user is reading or editing, based on similarities in the full text pattern.

Other projects such as Fab, Letizia, WebWatcher, Syskill and Webert [5,6,7,8] use agents that aim to derive a user context or profile by examining the navigational behavior of the user. These systems are referred to as adaptive information retrieval systems [9]. Based on derived user profiles, they customize web pages by rating or removing hyperlinks or by providing a separate panel with recommendations to other interesting documents.

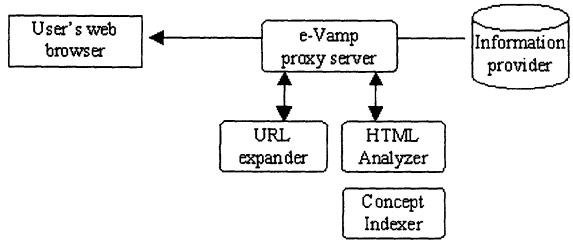

Figure 1 - Overview of the e-Vamp architecture. The browser will request a page from the web provider. In between, the e-Vamp server will enrich the page with hyperlinks that retrieve the relevant articles from a pre-indexed collection

Specialized search engines that cover a particular domain, provide another approach to context [10]. ResearchIndex [11] is a specialized search engine for scientific literature that uses the citations in papers as a context for a search and features special functions to extract information generally found in scientific publications.

Query formulation and expansion

Other projects, such as the CliniWeb [12] project strive to support the (novice) user with query formulation in stead of building a query context. In the CliniWeb project, different medical information sources have been indexed using a well-defined ontology. A query term is mapped to a concept in the ontology. Relations for the concept defined in the ontology are used to automatically derive a context that can be used for query expansion. The MedWeaver [13] project provides a thesaurus-based search engine that includes knowledge sources such as decision systems as well. In the Ariane project [14], a user query is automatically expanded with related terms from a thesaurus. Several relations from the Unified Medical Language System [15] semantic network are used to expand the query. In this paper, an approach is described that can annotate web-based information resources on-the-fly with thesaurus based concept information. Each recognized concept irrespective of the synonym used in the text can be clicked and a range of services has been implemented, ranging from viewing the thesaurus-definition of the concept to outlinking into protein and gene databases. One of the options is to perform an automated search in all connected information resources on the chosen concept, taking into account the context in which the term is mentioned in the text. The search is executed against thesaurus-based indexes of various different information sources. The query with context is directly available in the document as a pre-calculated contextual fingerprint and does not require the user to switch to another window to enter a query. The implemented approach is different from approaches where all the words in a document are used as search terms without assigning different specific weights to the various terms. In the approach presented here, a choice is offered between (a) context based on a surrounding paragraph or (b) on the full text.

Materials and Methods

The e-Vamp service has been based on the Collexis© indexing and retrieval software. This software has been based on the vector space model, but in stead of using the individual words in a text it uses concepts as the dimensions of the vector space. These concepts and their synonyms are defined in a thesaurus. The Collexis indexer uses this thesaurus to identify in a text the concepts and assign a relevance weight to each concept (similar to term frequency). The set of concepts found in a text with their associated relevance weight is called a fingerprint. This simple representation can be manipulated in different ways. For instance, concept fingerprints of documents written by an author can be accumulated into an author fingerprint [16].

Based on this technology a web service has been developed, called e-Vamp. It acts as a proxy server that resides in between the user's browser and the information provider (see figure 1). Whenever the user requests a web page via this e-Vamp proxy server, the following actions will be executed. The e-Vamp server retrieves the web page from the information provider. Next, it analyzes the web page and expands all hyperlink references from relative paths to absolute paths. Subsequently, all hyperlink references are relocated so that they also refer to the e-Vamp proxy server. The retrieved web page is fed to the analyzer. This analyzer parses the HTML layout and distills the pure text from the page. This text is fed to the concept indexer which will return a concept fingerprint to the e-Vamp service. The analyzer will use this information to insert new hyperlinks in the web page for each concept that has been found together with a context for the web page. The context is based on the concepts immediately surrounding the selected concepts and their relevance score. Finally, the analyzer will forward the HTML page to the user's web browser. The speed of the analyzer has been optimized such that the entire process described above can be handled within the normal response time of a web server. Obviously, the same service can be implemented with all sorts of texts as a starting point, not just with web pages.

Related information

A web page retrieved via the e-Vamp service is visually identical to the same web page when retrieved directly. However, if the mouse is moved over a word or phrase that has been recognized as belonging to a concept, it will light up. When pressing the left mouse button on such a word, the user gets a pop up menu with a number of possibilities (see Figure 2). The first distinguishing feature is that a definition of the concept is given in the pop up window. This is crucial when readers venture into new areas of exploration and do not necessarily know each concept, its definition and its common synonyms. A prominent example is given in figure 2. A user not familiar with the fact that Epstein-Barr virus (EBV) is the same concept as Human Herpes Virus 4 will be surprised, even if a query expander based on thesaurus information replaces the request for EBV automatically for the UMLS preferred term Human Herpes Virus 4.

With e-Vamp however, the concept remains as it was expressed in the text, but the query string inserted behind it contains the concept number and does not only lead to the definition, but also includes all known synonyms from the thesaurus.

Next to having all this information about the concept at their mouse-tip, users can search for similar documents (literature) based on a paragraph context or on the context from the whole text on that particular page. In addition to MedLine and on-line publications, several other resources have been added such as patents and news services. In this way, each e-Vamped text is instantly interactive with all pre-indexed data resources in the system, irrespective of their local query systems and preferred jargon. Last but not least, if a molecular concept (i.e. a gene or a protein) is also known in other databases, there is direct access to the records on the clicked concept by outlinking via the unique identifier of the molecule into that particular database.

As we recognize a trend from reading to browsing to consulting experts for quick reference and knowledge acquisition in new areas and disciplines the system also allows for an expert search in context, presently based on accumulated, computer-mined author-document relationships in MedLine. Using this feature

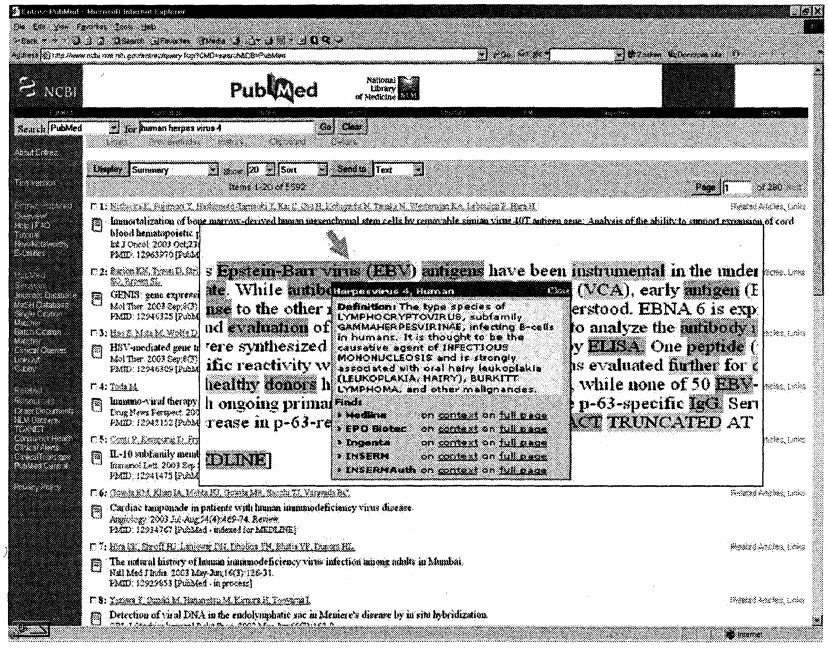

Figure 2 - An example of a page from PubMed e-Vamped. The pop up window whows the various outlinks for the Eptein Barr Virus concept as well as its definition. The concept is found independant of which synonym is used in the text

the reader can find authors (experts) that have published about the clicked concept in its context straight away for consultation.

Matching information

When selecting related documents or experts based on a paragraph, not only the clicked concept but also several of its neighbors in the text are included. In case the whole document is used, all concepts are included in the search. Since the technology allows for specifying a relevance score for each concept, the selected concept gets the maximum weight and its neighbors a decreasing weight depending on the distance to the selected concept. The surrounding concepts form the context for the selected concept.

The search concept fingerprint as contained in the e-vamped web page is sent to the Collexis matching service. This is a web-based service that uses an approximative approach to match the search concept fingerprint with a large number of content fingerprints (see Figure 3). The matching service computes the similarity between the fingerprint composed of a concept with its contextual concepts and all document fingerprints and returns those with the highest similarity value. It is important to note that using the context as a Boolean expression would often lead to zero hits. The approximative match tries to find those documents closest to the search fingerprint. The matching service will return the unique identifiers of those documents that match, together with meta-data that has been stored in the Metadata service.

Results

The matching service can use various collections of fingerprints. Based on the nature of an information source, concept fingerprints are stored in separate collections. The current environment contains a collection for full text fingerprints, MedLine abstract fingerprints, and patents. The matching service compares the search fingerprint with the fingerprints stored in the different collections. On a regular Pentium 4 server, 2.5 million fingerprints can be compared in 300 ms. This includes the retrieval of the meta-data for the best matching document fingerprints.

Technical evaluation

The e-vamp service analyses full web pages and rewrites them with new inserted hyperlink references. A first technical evaluation criterion was that structure and layout of the original page should not be affected by e-vamp. After an extensive technical evaluation of a number of web pages no visual difference between original and e-Vamped page was found. A second aspect of the technical evaluation dealt with the alteration of original hyperlink references with the address of the proxy server. A modified hyperlink will cause any reference to be e-Vamped as well. We can produce the relocation for static defined hyperlink references, but excluded those references that are dynamically composed by a script.

User evaluation

The current version of the e-Vamp service has been informally evaluated by various professionals.

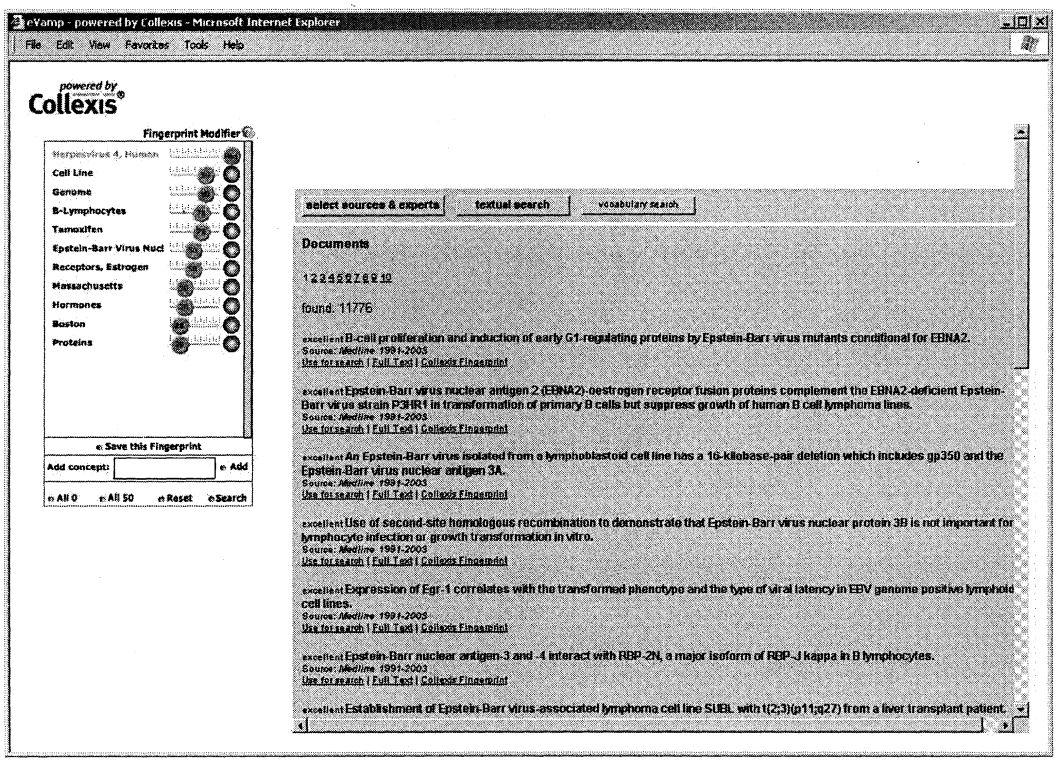

Figure 3 - The contextual search as initiated by e-Vamp. In the left pane the concepts are shown (with diminishing weights) that are used for searching. In the right pane one sees the approximative search results

The first results are very positive and stimulated us to design a number of real world use scenarios that can be used for a more formal user evaluation. A first use scenario that will be implemented is the automatic annotation of online clinical guidelines and protocols. A second test case will be the automatic annotation of the pages describing genes or proteins in the online. A small-scale user evaluation in our department yielded very positive reactions. The system is highly intuitive as browsing can be started on a mouse click from each concept in the text under study. The pop up of the definition was seen as highly useful for less well established concepts. A formal user evaluation of this tool will be done in the context of the developed EMBO e-Biosci system

Conclusion

On the fly parsing of web pages and the synthesis of new hyperlinks is now technically feasible and at a speed that allows practical implementation. This paper describes the technology and the first prototype with some considerations on the basic philosophy behind this new approach to browsing of massive amounts of related resources on the Web. Having a page on the screen as a fully "connected" set of concepts in context rather than as an electronic version of a flat text will dramatically improve web publishing advantages. The inclusion of a concept context in a related search from within a document is likely to significantly improve navigation through associated resources, in particular when ambiguity in concept names plays an important role in the subject area, which is the case in most life sciences area's but mainly in current high-interest research topics such as genomics and proteomics. The possibility to link out during the browsing process to a categorization of content using a concept context and an approximative search engine gives users a completely new tool that diminishes the difference between browsing and searching.

References

[1] Salton G, McGill MJ. *Introduction to Modern Information Retrieval*, 1983.

[2] Glover EJ, Lawrenc S, Gordon MD, Birmingham WP, Giles CL. Web search – your way. *Comm of the ACM*, 2000:235-42.

[3] Budzik J, Hammond KJ. User interactions with everyday applications as context for just-in-time information access. In *Proc 2000 Int Conf on Intell User Interf*, New Orleans, Louisiana, 2000: 44-51.

[4] www.kejin.com

[5] Balabanovic M. An adaptive web recommendation service. In: *Proc of the First Int Conf on Autonomous Agents*, 1997:378-385.

[6] Liebermann H. Letizia: An agent that assist web browsing. In: 1995 *Int Joint Conf on Art Intell*, Montreal CA, 1995;22:924-929.

[7] Armstrong D, Freitag D, Joachims T, Mitchell T. *WebWatcher: A learning apprentice for the World Wide Web.* 1995.

[8] Pazzani M, Muramatsu J, Billsus D. Syskill & Webert: Identifying interesting web sites. In: *Proc National Conf on Artif Intell*, AAAI, 1996:54-61.

[9] Lawrence S. Context in Web Search. *IEEE Data Engineering Bulletin*, 2000;23(3):25-32.

[10]www.completeplanet.com, www.invisibleweb.com

[11]Lawrence S, Giles CL, Bollacker K. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67-71, 1999.

[12]Hersh WR, Brown KE, Donohoe LC, et al. CliniWeb: managing clinical information on the World Wide Web. *J Am Med Inform Assoc* 1996;3:273-80.

[13]Detmer WM, Barnett GO, Hersh WR. MedWeaver: integrating decision support, literature searching, and Web exploration using the UMLS MetaThesaurus. In: *Proc of the AMIA Ann Fall Symp.* 1997:490-4.

[14]Joubert M, Fieschi M, Robert JJ, Volot F, Fieschi D. UMLS-based conceptual queries to biomedical information databases: an overview of the project ARIANE. *J Am Med Inform Assoc.* 1998;5:52-61.

[15]Nelson SJ, et al. A Semantic Normal Form for Clinical Drugs in the UMLS: Early Experiences with the VANDF. In: Kohane, Issac S., editor. Bio*medical Informatics: One Discipline. *Proc of the AMIA Ann Symp.* 2002:557-61.

[16]Van Mulligen EM. Diwersy M, Schmidt M, Buurman H, Mons B. Facilitating networks of information. *Proc of the AMIA Ann Symp.* 2000:868-72.

Address for correspondence:

Erik M. van Mulligen PhD

Dept of Medical Informatics,

Erasmus Medical Center Rotterdam

P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands

e.vanmulligen@erasmusmc.nl