Placing Search in Context: The Concept Revisited

Lev Finkelstein, Evgeniy Gabrilovich[1], Yossi Matias, Ehud Rivlin,
Zach Solan, Gadi Wolfman and Eytan Ruppin
Zapper Technologies Inc.
3 Azrieli Center, Tel Aviv 67023, Israel
+972-3-6949222
{lev,gabr,yossi,ehud,zach,gadi,eytan}@zapper.com

ABSTRACT

We describe a new paradigm for performing search in context. In the IntelliZap system we developed, search is initiated from a text query marked by the user in a document she views, and is guided by the text surrounding the marked query in that document ("the context"). The context-guided information retrieval process involves semantic keyword extraction and clustering to automatically generate new, augmented queries. The latter are submitted to a host of general and domain-specific search engines. The results are then semantically reranked, again, using context. It is our belief that letting context guide the search provides a better match to the user's current needs than just relying on the user's fixed personal profile. Our results show that using context to guide search effectively offers even inexperienced users an advanced search tool on the Web.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, query formulation, search process*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *dictionaries, linguistic processing, thesauruses*; I.7.5 [**Document and Text Processing**]: Document Capture – *document analysis*.

General Terms

Algorithms, Performance.

Keywords

Search, Context, Semantic Processing, Invisible Web, Statistical Natural Language Processing.

1. INTRODUCTION

Given the constantly increasing information overflow of the digital age, the importance of information retrieval has become critical. Web search is today one of the most challenging problems of the Internet, striving at providing users with search results most relevant to their information needs.

Internet search engines evolved through several generations since their inception in 1994, progressing from simple keyword matching to techniques such as link analysis and relevance feedback (achieved through refinement questions or accumulated personalization information) [11]. Search engines have now entered their third generation, and current research efforts continue to be aimed at increasing coverage and relevance.

A large number of recently proposed search enhancement tools have utilized the notion of *context*, making it one of the most abused terms in the field, referring to a diverse range of ideas from domain-specific search engines to personalization. We present here a novel search approach that interprets context in its most natural setting, namely, a body of words surrounding a user-selected phrase. We postulate that a large fraction of searches originate while users are reading documents[2] on their computers, and require further information about a particular word or phrase. Hence, the basic premise underlying our approach is that searches should be processed *in the context of the information surrounding them*, allowing more accurate search results that better reflect the user's actual intensions. For example, a search for the word "Jaguar" should return car-related information if performed from a document on the motoring industry, and should return animal-related information if performed from an Internet website about endangered wildlife. Guiding user's search by the context surrounding the text eliminates possible semantic ambiguity and vagueness.

Our system (named IntelliZap) is based on the client-server paradigm, where a client application running on user's computer captures the context around the text highlighted by the user. The server-based algorithms analyze the context, selecting most important context words and performing word sense disambiguation, and then prepare a set of augmented queries for subsequent search. The technology also enables the user to modify the extent to which context guides a specific search, by modifying the amount of context considered. Queries resulting from context analysis are dispatched to a number of search engines, performing meta-searching. When the context can be reliably classified to a predefined set of domains (such as health, sport or finance), additional queries are dispatched to search engines specializing in this domain. This step can be viewed as referring to the *Invisible*

WWW10, May 1-5, 2001, Hong Kong.
ACM 1-58113-348-0/01/0005.

[1] Corresponding author (email: gabr@zapper.com).

[2] Such documents can be in a variety of formats (MS Word DOC, HTML or plain text to name but a few), and either online (residing on the Internet) or offline (residing on a local machine).

Web, as some of the target domain-specific engines may constitute front-ends to databases that are not otherwise indexed by conventional search engines. A dedicated *reranking* module ultimately reorders the results received from all the engines, according to semantic proximity between their summaries and the original context. To this end we use a semantic metric that given a pair of words or phrases returns a (normalized) score reflecting the degree to which their meanings are related.

The significance of the new *context-based* approach lies in the greatly improved relevance of search results. We achieve this by applying natural language processing techniques to the captured context in order to guide the subsequent search for user-selected text. Existing approaches either analyze the entire document the user is working on, or ask the user to supply a category restriction along with search keywords. As opposed to these, the proposed method analyzes the context in the immediate vicinity of the focus text. This allows analyzing just the right amount of background information, without running over the more distant (and less related) topics in the source document. The method also allows collecting contextual information without conducting an explicit dialog with the user.

This paper is organized as follows. The next section reviews related work. Section 3 presents the various features of our context-based search system, explaining how several individual algorithms work in concert to improve the relevance of the search. Section 4 discusses the experimental results. Finally, Section 5 concludes the paper and suggests further research directions.

2. RELATED WORK

Using context for search is not a new idea. A number of existing information retrieval systems utilize the notion of context to some extent. The problem is, however, that everyone defines *context* a little differently.

Lawrence [8] contains an elaborate review of using context in Web search. Explicit context information can be supplied to a search engine in the form of a category restriction[3]. Such a category may considerably disambiguate a query and thus focus the results. For instance, given the search term "jaguar", possible categories are "fauna" or "cars". Inquirus-2 project [7] specifically requests context information in this way.

In contrast to this approach, other tools infer context information automatically by analyzing whole documents displayed on users' screens. The Watson project [2] attaches this background information to explicit user queries, while tools like Kenjin[4] automatically suggest Web sites related[5] to the document being worked upon. Such tools encounter difficulties when documents are long and discuss a variety of topics – as the data collected from the entire document reflects all the topics covered, it might not be particularly relevant to the user's current focus (be it an explicit query in the former case, or simply the active part of the document in the latter). The main difference between such tools and our IntelliZap is that the latter analyzes the context *in the immediate vicinity* of the user-selected text, thus making the context coherent and focused around a single topic. In the other end of the spectrum, tools like GuruNet (now Atomica[6]) offer database lookup directly from reference sources (dictionaries, encyclopedias etc.). Such tools offer only a limited usage of text, without deep semantic analysis of the enclosing context.

There is a family of tools that interpret the notion of context as a set of previous information requests originated by a user. Defined this way, context search becomes *personalization*, and tools in this category keep track of user's previous queries and/or documents viewed. SearchPad [1] recognizes that many advanced users perform several searches concurrently, and tracks search process over time. This extension to search engines keeps track of "search context" by following the different search sessions and collecting "useful queries and promising results links" [1].

Other ways of incorporating context into search include the usage of domain-specific rather than general-purpose engines [8]. Databases which belong to the Invisible Web (i.e., whose contents are not indexed by conventional search engines) may be particularly useful as they might contain vast amounts of information within their narrow domain. IntelliZap pursues a similar approach by classifying the topic of the query context, and targeting search engines specializing in the corresponding domain. Note that this way the selection of specialized search engines is performed automatically.

Yet another interpretation of *context* belongs to the realm of *link analysis* [12, 13]. In the quest to expand the coverage, some engines intentionally limit the number of sites they index to make the retrieval efficient, but can still yield "unindexed" sites in search results. This is achieved by analyzing the *context of links* pointing at these sites, thus deducing information about the contents of the target. Google and Inktomi[7], among others, employ this technique. Another context-related feature of Google shows up in its search-dependent result summaries. A typical Google summary contains an excerpt from the Web page where the search terms are shown highlighted *in the context of this page* [14].

In contradistinction to this variety of interpretations of context usage in search, our approach focuses on using the context in its most natural sense – that of the text surrounding the marked query, providing local semantic consistency for its interpretation.

3. THE CONTEXT-BASED SEARCH (CBS) SYSTEM

Current approaches to information retrieval over the Web are based on a scenario in which the user enters a query to a search engine. The search engine then retrieves a set of ordered documents that best match the user's query. We propose an approach that changes the basic settings of the search scene by using the context of the query as an additional input. In this

[3] The target engine must obviously support a mechanism for search restriction, so that a category constitutes an integral part of the query.

[4] www.kenjin.com

[5] Note that Kenjin provides related links as opposed to performing conventional search.

[6] www.atomica.com

[7] www.google.com and www.inktomi.com, respectively.

scenario, when a user marks a text in a document and submits it for search, the system captures the context surrounding the text, and utilizes it to yield more focused results. The context may include the sentence containing the query word or phrase, a few sentences surrounding the query term, the paragraph in which it resides, or even the whole document.

Using the context for superior search focus constitutes a considerable algorithmic challenge. One needs to find ways to extract the right amount of context which best optimizes the information retrieved, as well as devise adequate ways to use the context extracted for focusing the response to the original query.

Apparently, the simplest way to do so is to concatenate the selected text and the captured context into a uniform query, which can be sent to search engines. However, this approach does not perform satisfactorily. First, most search engines cannot handle large chunks of natural language text. Second, blindly concatenating long context with relatively short text might cause search results to become greatly unfocused.

In what follows, we describe two fundamental approaches to building augmented queries using contextual information. The next section outlines two relatively simple heuristic techniques. The section that follows presents more powerful query augmentation methods, which identify the most important context words to guide the subsequent information retrieval process.

3.1 Heuristic Query Augmentation System

We evaluated two classes of heuristics. Each of them receives as input the marked text and its context captured from some application.

1. *Establishing optimal context length as a function of the length of the text phrase and individual word frequencies.*

   We discovered that the context length should be commensurate with the text length, and with the relative frequencies of text words. That is, the more frequent a text word is, the less information it likely carries and the larger context should be supplied to focus the search.

2. *Relative weighting of the text and context in the augmented query.*

   Weighting is performed by word duplication, requiring certain words (estimated to carry substantial amount of information) to appear in titles of retrieved documents, as well as by using Boolean operators (AND, OR, NEAR etc.). In general, such operators emphasize the marked text phrase in the augmented query built, and make the weight of context words in the augmented query a monotonic function of their proximity to the text phrase.

Figure 1 shows an example of query augmentation using the heuristics. The heuristic approach yielded a certain improvement in the relevance of the information retrieved to the user's interests, and served as an encouraging example of the potential of utilizing context to guide information search. Yet, further significant improvement is possible by pursuing a more general approach, which utilizes background linguistic information and does not depend on the specific syntax of search engines input, i.e., does not require the knowledge of the specific operators recognized by each potential target search engine.

3.2 Linguistic CBS System

3.2.1 Overview

We have developed a system called IntelliZap[8] that performs context search from documents on users computers. When viewing a document, the user marks a word or phrase (referred to as *text*) to be submitted to the IntelliZap service (in the example of Figure 2 the marked text is the word "jaguar"). The client application automatically captures the *context* surrounding the marked text, and submits both the text and the context to server-based processing algorithms.

Figure 2 shows a screen shot with the software client invoked on a user document, and Figure 3 demonstrates a part of the results page. Observe that the top part of the results page repeats the user-selected text *in the original context* (only part of which is displayed, as the actually captured context may be quite large).

Input
Text: vaccines
Context: Flu experts say the answer's very simple: vaccination distribution is a private endeavor. The federal government makes sure that vaccines are safe and effective, but doesn't have much involvement in distribution.

Output
Augmented query: `+vaccines title:distribution Flu Flu experts vaccination distribution private endeavor federal government vaccines safe effective distribution`

Figure 1. An example of query augmentation using heuristic techniques

[8] The IntelliZap client application may be obtained from www.zapper.com. The Web site also features a Web-based IntelliZap, which does not require client download, but rather allows to copy-and-paste both search terms and context into appropriate fields of an HTML form. The latter feature is available at http://www.zapper.com/intellizap/intellizap.html.

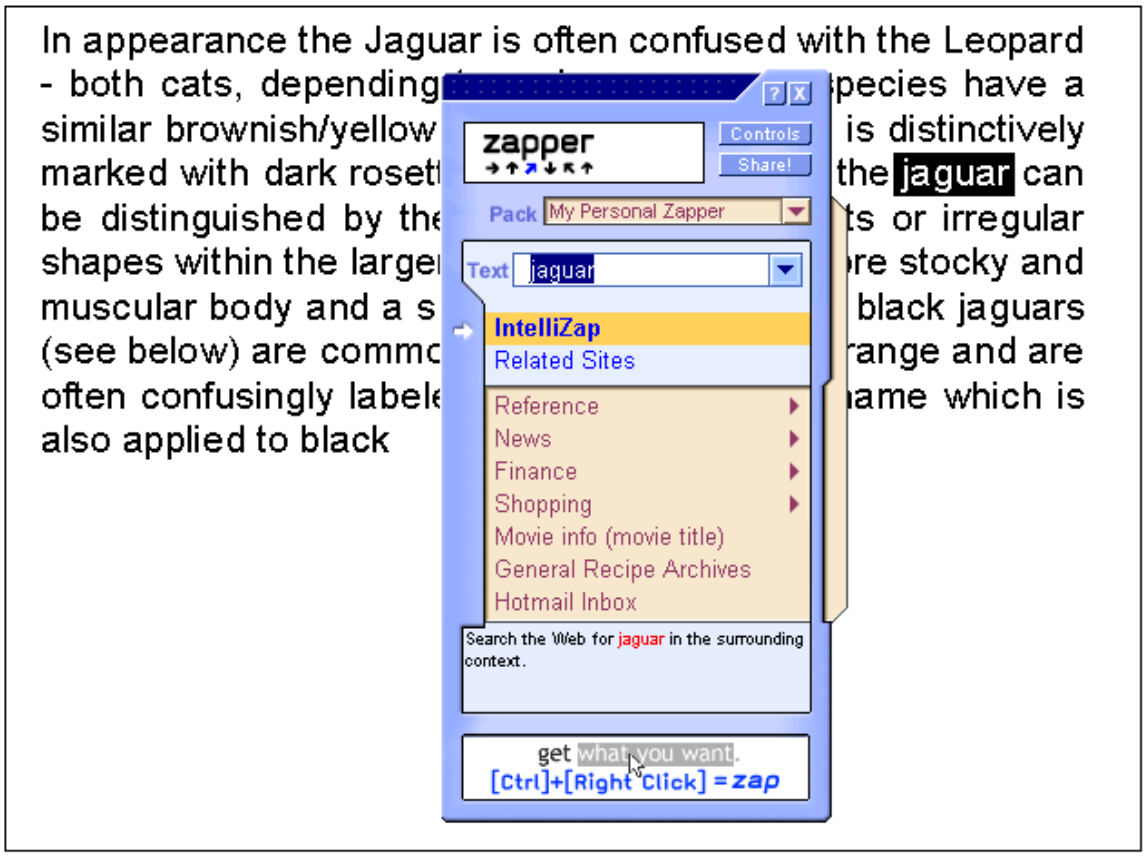

Figure 2. IntelliZap client invocation on a document

Figure 3. IntelliZap search results

3.2.2 The Core Semantic Network

The core of IntelliZap technology is a semantic network which was designed to provide a metric for measuring distances between pairs of words. The semantic network is implemented using a vector-based approach, where each word is represented as a vector in multi-dimensional space. To assign each word a vector representation, we first identified 27 knowledge domains (such as computers, business and entertainment) roughly partitioning the whole variety of topics. We then sampled a large set of documents in these domains[9] on the Internet. Word vectors[10] were obtained by recording the frequencies of each word in each knowledge domain. This way each domain can be viewed as an axis in the multi-dimensional space. The distance measure between word vectors is computed using a correlation-based metric. Although such a metric does not possess all the distance properties (observe that the triangle inequality does not hold), it has strong intuitive grounds: if two words are used in different domains in a similar way, these words are most probably semantically related.

We further enhance the statistically based semantic network described above using linguistic information, available through the WordNet electronic dictionary [9]. Since some relations between words (like hypernymy/hyponymy and meronymy/holonymy) cannot be captured using purely statistical data, we use WordNet dictionary to correct the correlation metric. A WordNet-based metric was developed using an information content criterion similar to [10], and the final metric was chosen as a linear combination between the vector-based correlation metric and the WordNet-based metric.

Currently, our semantic network is defined for the English language, though the technology can be adapted for other languages with minimal effort. This would require training the network using textual data for the desired target language, properly partitioned into domains. Linguistic information can be added subject to availability of adequate tools for the target language (e.g., EuroWordNet [5] for European languages or EDR [15] for Japanese).

The IntelliZap system has three main components based on the semantic network:

1. Extracting keywords from the captured text and context.
2. High-level classification of the query to a small set of predefined domains.
3. Reranking the results obtained from different search engines.

Figure 4 gives a schematic overview of the IntelliZap algorithm. The following sections explain the individual components listed above.

3.2.3 Keyword Extraction Algorithm

The algorithm utilizes the semantic network to extract keywords from the context surrounding the user-selected text. These keywords are added to the text to form an augmented query, leading to context-guided information retrieval.

The algorithm for keyword extraction belongs to a family of clustering algorithms. However, a straightforward application of such algorithms (e.g., K-means [4, 6]) is not feasible due to a large amount of noise and a small amount of information available: usually we have about 50 context words represented in 27-dimensional space, which makes the clustering problem very difficult. In order to overcome this problem, we developed a special-purpose clustering algorithm, which performs recurrent clustering analysis, and refines the results statistically. For a typical query of 50 words (one to three words in the text, and the rest in the context), the keyword extraction algorithm usually returns three or four clusters, which correspond to different aspects of the query. Cluster-specific queries are built by combining text words with several most important keywords of each cluster. Responding to such queries, search engines yield results covering most of the semantic aspects of the original query, while the reranking algorithm filters out irrelevant results.

3.2.4 Search Engine Selection

The queries created as explained above are dispatched to a number of general-purpose search engines. In addition, we attempt to classify the captured context in order to select domain-specific search engines that stand a good chance of providing more specialized results. The classification algorithm classifies the context to a limited number of high-level domains[11] (e.g., medicine or law). A probabilistic analysis determines the amount of similarity between the domain signatures and the query context. The *a priori* assignment of search engines to domains is performed offline.

Some of the search engines (such as AltaVista[12]) allow limiting the search to a specific category. In such cases, categorizing the query in order to further constrain the search usually yields superior results.

3.2.5 Reranking

After queries are sent to the targeted search engines, a relatively long list of results is obtained. Each search engine orders the results using its proprietary ranking algorithm, which can be based on word frequency (inverse document frequency), link analysis, popularity data, priority listing, etc. Therefore, it is necessary to devise an algorithm which would allow us to combine the results of different engines and put the most relevant ones first.

[9] Approximately 10,000 documents have been sampled in each domain.

[10] Each word vector has 27 dimensions, as the number of different domains.

[11] Currently, nine domains are defined, each of which is mapped to two or three search engines.

[12] www.altavista.com

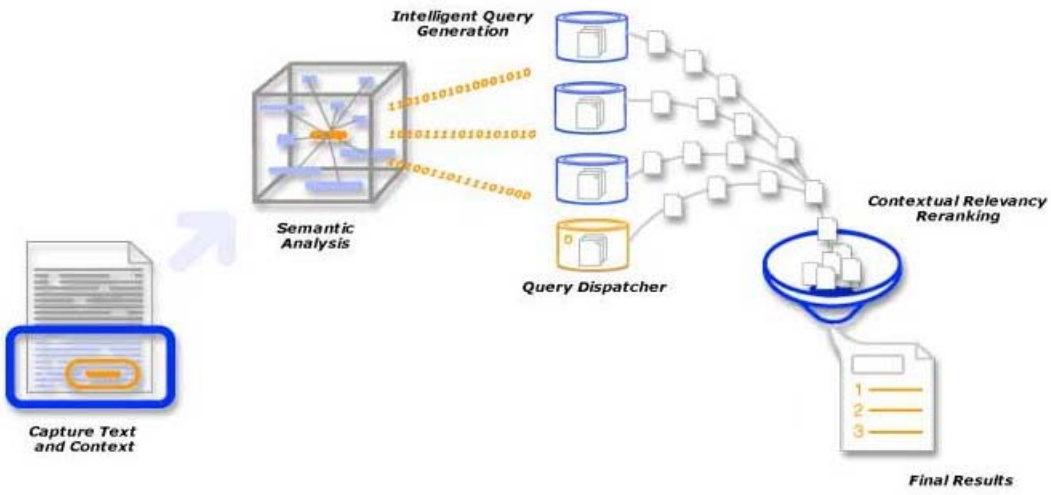

Figure 4. IntelliZap algorithm overview: information and processing flow (from left to right)

At first, this problem may seem misleadingly simple – after all, humans usually select relevant links by quickly scanning the list of results summaries. Automating such an analysis can, however, be very demanding. To this end, we make use of the semantic network again, in order to estimate the relatedness of search results to the query context.

Our reranking algorithm reorders the merged list of results by comparing them semantically with both text and context. The algorithm computes semantic distances between the words of results titles and summaries on the one hand, and the words of text and context on the other hand. An important feature of the algorithm is that the distances computed between text (context) and summaries are *not* symmetric. As we observed in experiments, user usually welcome results whose summaries are *more general* than the query, but tend to ignore results whose summaries are *more specific* than the query. Each summary is given a score based on the distances computed between text and summary, summary and text, context and summary, and summary and context. Search results are sorted in decreasing order of their summary scores, and the newly built results list is displayed to the user.

4. EXPERIMENTAL RESULTS

In this section we discuss a series of experiments conducted on the IntelliZap system. The results achieved allow us to claim that using the context effectively provides even inexperienced users with advanced abilities of searching the Web.

4.1 Context vs. Keywords: A Quantitative Measure

A survey conducted by the NEC Research Institute shows that about 70% of Web users typically use only a single keyword or search term [3]. The survey further shows that even among the staff of the NEC Research Institute itself, about 50% of users use one keyword, additional 30% – two keywords, about 15% – three keywords, while only 5% of users employ four keywords or more.

The goal of the experiment described below was to determine what number of keywords in a keyword-based search engine is equivalent to using the context with our IntelliZap system.

Twenty-two subjects recruited by an external agency participated in this study. Conditions for participation included at least minimal acquaintance with the Internet and high level of English command. Each subject was presented with three short texts and was asked to find (in three separate stages of the test) information relevant to the text using IntelliZap and each of the following search engines: Google, Yahoo!, AltaVista, and Northern Light[13]. The subjects were told that the study compares the utility of a variety of engines. At no point were they informed that the comparison between IntelliZap specifically and the other engines was the focus of the study. The subjects were asked to search for relevant information using one, two and three keywords using each of the search engines. The instructions for using IntelliZap remained the same through all stages – to capture any word or phrase from the text, as the users deemed appropriate. Relevancy[14] was rated for the first ten results returned. The rating system was defined as follows: 0 for irrelevant results, 0.5 for results relevant only to the general subject of the text, and 1 for results relevant to the specific subject of the text. Dead links and results in languages other than English were assigned the score of 0. Figures 5, 6 and 7 show the results for one, two, and three keyword queries, respectively. The non-monotonic behavior of the number of relevant results among the stages is due to the usage of different texts (as explained above).

As evident, using the context efficiently enables IntelliZap to outperform other engines even when the latter are probed with three-keyword queries.

[13] www.google.com, www.yahoo.com, www.altavista.com, and www.northernlight.com, respectively.

[14] The notion of *relevancy* was obviously subjectively interpreted by each tester. Here we report the cumulative results for all the participants of the experiment.

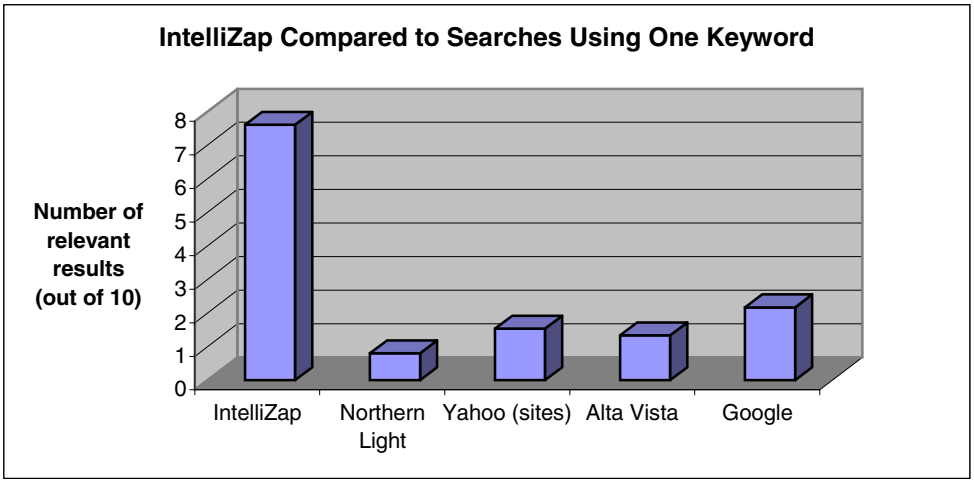

Figure 5.

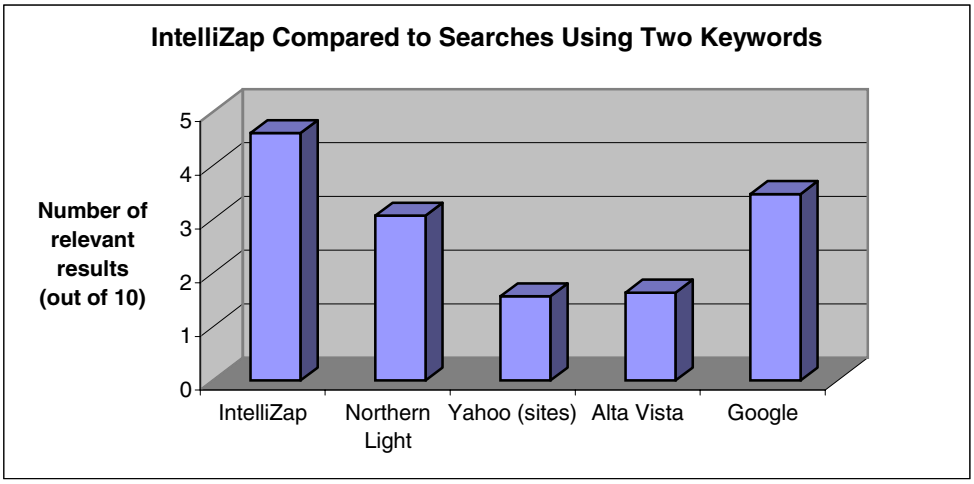

Figure 6.

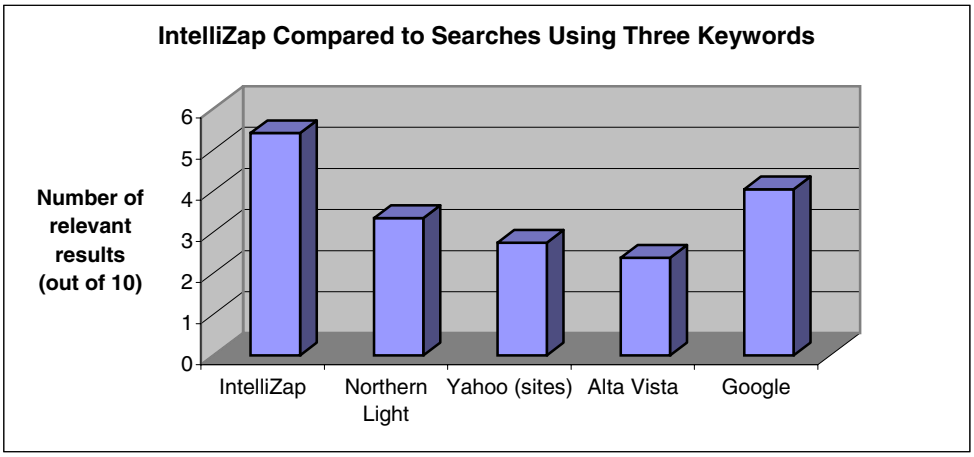

Figure 7.

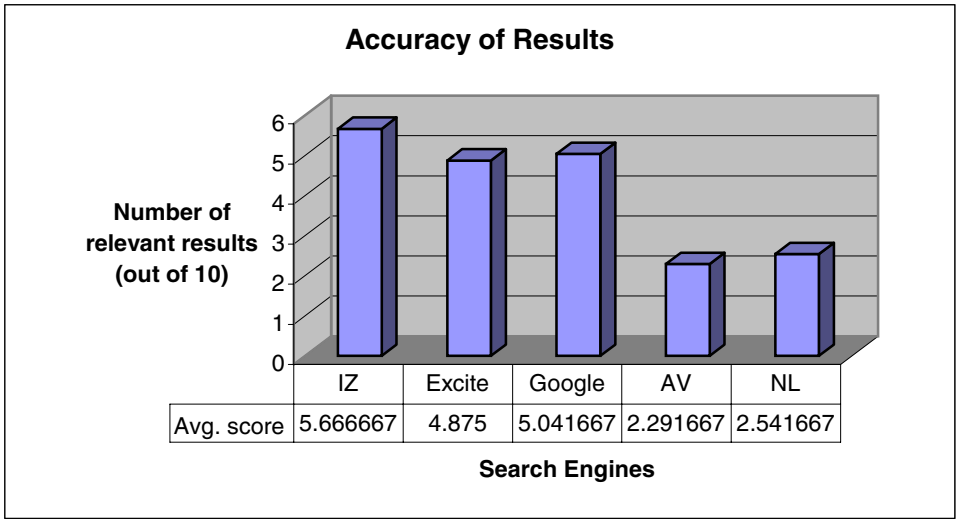

Figure 8. IntelliZap vs. other search engines: accuracy of results

4.2 IntelliZap vs. Other Search Engines: An Unconstrained Example

In order to validate the IntelliZap performance, we compared it with a number of major search engines: Google, Excite, AltaVista, and Northern Light[15]. Twelve subjects recruited by an external agency were tested. As before, the subjects were required to have some acquaintance with the Internet and high level of English command. At no point throughout the study were the subjects explicitly informed that the comparison between IntelliZap specifically and the other engines was the focus of the study.

Each subject was presented with five randomly selected short texts. For each text the subject was asked to conduct *one* search in order to find information relevant to the text using a randomly assigned search engine. The subjects were given no instructions or limitations regarding how to search. This is because the aim of this part of the test was to compare IntelliZap to other search engines when users employed their natural search strategies. In particular, the users were allowed to use boolean operators and other advanced search features as they saw fit. The IntelliZap system used in this experiment utilized Google, Excite, Infoseek[16] (currently GO network search) and Raging Search[17] as underlying general-purpose engines. A number of domain specific search engines (such as WebMD and FindLaw[18]) were also used in cases when the high-level classification succeeded in classifying the domain of the query. The subjects were required to estimate the quality of search by counting the number of relevant links in the first ten results returned by each engine. The relevancy rating system was identical to the one described in the previous experiment.

As can be seen from the comparison chart in Figure 8, IntelliZap outperforms the rest of search engines. Note that the above test measures only the precision of search, as it is very difficult to measure the recall rate when operating Web search engines. However, the precision rate appears to be highly correlated with the user satisfaction from search results.

4.3 Response Time

In the client-server architecture of IntelliZap, client-captured text and context are sent for processing to the server. Server-side processing includes query preparation based on context analysis, query dispatch, merging of search results, and finally delivering the top reranked results to the user. The cumulative server-side processing time per user query is less than 200 milliseconds, measured on a Pentium III 600 MHz processor. In contrast to the conventional scenario, in which users access search engines directly, our scheme involves two connection links, namely, between the user and the server, and between the server and search engines (that are contacted in parallel). Thanks to the high speed Internet connection of the server, the proposed scheme delivers the results to the end user in less than 10 seconds.

5. DISCUSSION

This paper describes a novel algorithm and system for processing queries in their context. Our approach caters to the growing need of users to search directly from items of interest they encounter in the documents they view. Using the context surrounding the marked queries, the system enables even inexperienced web searchers to obtain satisfactory results. This is done by autonomously generating augmented queries, and by autonomously selecting relevant search engine sites to which the queries are targeted. The experimental results we have presented testify to the very significant potential of the approach.

Our work opens up a new and promising avenue for information retrieval, but much future work could and should be done to carry it further. Among the rest, context should be utilized to expand the

[15] www.google.com, www.excite.com, www.altavista.com, www.northernlight.com, respectively.

[16] www.go.com

[17] www.raging.com

[18] www.webmd.com and www.findlaw.com, respectively.

augmented queries in a disambiguated manner. In fact, this disambiguation process could be used to concomitantly determine the extent of the context which is most relevant for processing the specific query in hand. More work could be done on specifically tailoring the generic approach shown here for maximizing the context-guided capabilities of individual search engines (applying our algorithm in such a manner to one of the leading major search engines has provided very encouraging results).

Interestingly, we find a seemingly paradoxical effect in applying our context-guided search to various search engines: the better the engine is, the more it can benefit from such context-dependent augmentation. This probably occurs because such engines are better geared up to process the semantically focused augmented queries with higher resolution, and respond more sharply and precisely to such well crafted queries. In summary, harnessing context to guide search from documents offers a new and promising way to focus search and counteract the "flood of information" so characteristic of search on the World Wide Web.

6. REFERENCES

[1] K. Bharat. SearchPad: Explicit Capture of Search Context to Support Web Search. In *Proceedings of the 9th International World Wide Web Conference, WWW9*, Amsterdam, May 2000.

[2] J. Budzik and K.J. Hammond. User interactions with everyday applications as context for just-in-time information access. In *Proceedings of the 2000 International Conference on Intelligent User Interfaces*, New Orleans, Louisiana, 2000. ACM Press.

[3] D. Butler. Souped-up search engines. *Nature*, Vol. 405, pp.112-115, May 2000

[4] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* New York: John Wiley and Sons, 1973.

[5] EuroWordNet. http://www.hum.uva.nl/~ewn/

[6] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* San Diego, CA: Academic Press, 1990.

[7] E. Glover et al. Architecture of a meta search engine that supports user information needs. In *8th International Conference on Information and Knowledge Management, CIKM 99*, pp. 210-216, Kansas City, Missouri, November 1999.

[8] S. Lawrence. Context in Web Search. *Data Engineering*, IEEE Computer Society, Vol. 23, No. 3, pp. 25-32, September 2000. http://www.research.microsoft.com/research/db/debull/A00sept/lawrence.ps

[9] G. Miller et al. WordNet – a Lexical Database for English. http://www.cogsci.princeton.edu/~wn

[10] P. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, Vol. 11, pp. 95-130, 1999.

[11] C. Sherman. Inktomi inside. http://websearch.about.com/internet/websearch/library/weekly/aa041900a.htm

[12] C. Sherman. Link Building Strategies http://websearch.about.com/internet/websearch/library/weekly/aa082300a.htm

[13] D. Sullivan. Numbers, Numbers – But What Do They Mean? *The Search Engine Report*, March 3, 2000. http://searchenginewatch.com/sereport/00/03-numbers.html

[14] The Basics of Google Search. http://www.google.com/help/basics.html

[15] T. Yokoi. The EDR Electronic Dictionary. *Communications of the ACM*, Vol. 38, No. 11, pp. 42-44, November 1995.