(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number : **0 597 630 A1**

# (12) EUROPEAN PATENT APPLICATION

(21) Application number : **93308829.6**

(51) Int. Cl.[5] : **G06F 15/403,** G06F 15/20

(22) Date of filing : **04.11.93**

(30) Priority : **04.11.92 US 970718**

(43) Date of publication of application :
**18.05.94 Bulletin 94/20**

(84) Designated Contracting States :
**AT BE CH DE DK ES FR GB GR IE IT LI LU MC NL PT SE**

(71) Applicant : **CONQUEST SOFTWARE INC.**
**9700 Patuxent Woods Drive, Suite 140**
**Columbia, Maryland MD-21046 (US)**

(72) Inventor : **Addison,Edwin R. Conquest Software Inc.**
**9700 Patuxent Woods Drive,Suite 140, Columbia,Maryland MD-21046 (US)**
Inventor : **Blair,Arden S. Conquest Software Inc.**
**9700 Patuxent Woods Drive,Suite 140, Columbia,Maryland MD-21046 (US)**
Inventor : **Nelson,Paul E. Conquest Software Inc.**
**9700 Patuxent Woods Drive,Suite 140, Columbia,Maryland MD-21046 (US)**
Inventor : **Schwartz,Thomas Conquest Software Inc.**
**9700 Patuxent Woods Drive,Suite 140 Columbia,Maryland MD-21046 (US)**

(74) Representative : **Goodman, Christopher**
**Eric Potter & Clarkson St. Mary's Court St. Mary's Gate**
**Nottingham NG1 1LE (GB)**

## (54) Method for resolution of natural-language queries against full-text databases.

(57) The method of the present invention combines concept searching, document ranking, high speed and efficiency, browsing capabilities, "intelligent" hypertext, document routing, and summarization (machine abstracting) in an easy-to-use implementation. The method of the present invention also offers Boolean and statistical query options. The method of the present invention is based upon "concept indexing" (an index of "word senses" rather than just words.) It builds its concept index from a "semantic network" of word relationships with word definitions drawn from one or more standard human-language dictionaries. During query, users may select the meaning of a word from the dictionary during query construction, or may allow the method to disambiguate words based on semantic and statistical evidence of meaning. This results in a measurable improvement in precision and recall. Results of searching are retrieved and displayed in ranked order. The ranking process is more sophisticated than prior art systems providing ranking because it takes linguistics and concepts, as well as statistics into account.

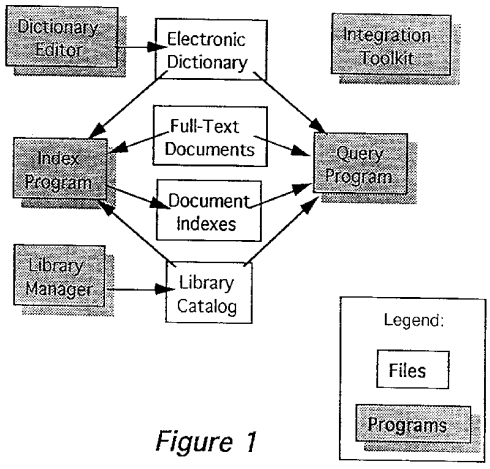


*Figure 1*

EP 0 597 630 A1

## Field of the Invention

The present invention is a method for computer-based information retrieval. Specifically, the method of the present invention comprises a computer-implemented text retrieval and management system. The present invention offers four advances in the art of computer-based text retrieval. First, querying is simple. Queries may be expressed in plain English (or in another suitable human language). Second, searching for "concepts" has been found to be more accurate than Boolean, keyword or statistical searching as practiced in the prior art. Third, the method of the present invention is more efficient than sophisticated text retrieval methods of the prior art. It is faster (in equivalent applications), and features recall in excess of 80%, as compared to recall of less than 30% for Boolean systems, and approximately 50% for statistical methods of the prior art. Finally, the method of the present invention manages the entire research process for a user.

## Background of the Invention

While there are dozens of information retrieval software systems commercially available, most of them are based on older Boolean search technology. A few are based on statistical search techniques which have proven to be somewhat better. But, to break the barrier to access to relevant information and to put this information in the hands of end users at the desktop requires search software that is intuitive, easy to use, accurate, concept oriented, and needs a minimum investment of time by the user. The following distinctive features and benefits delineate these significant aspects of the method of the present invention.

To date, there have been three major classes of text retrieval systems:

- Keyword or Boolean systems that are based on exact word matching
- Statistical systems that search for documents similar to a collection of words
- Concept based systems that use knowledge to enhance statistical systems

Keyword or Boolean systems dominate the market. These systems are difficult to use and perform poorly (typically 20% recall for isolated queries). They have succeeded only because of the assistance of human experts trained to paraphrase queries many different ways and to take the time to humanly eliminate the bad hits. While statistical search systems have increased performance to near 50% recall, trained search expertise is still needed to formulate queries in several ways to conduct an adequate search.

A concept based search system further closes the performance gap by adding knowledge to the system. To date, there is no standard way to add this knowledge. There are very few concept based search systems available and those that exist require intensive manual building of the underlying knowledge base.

The next logical direction for improvement in text retrieval is its use of Natural Language Processing (NLP). While there are some experimental systems in government development programs, most of those prototypes have been only useful in narrow subject areas, they run slowly, and they are incomplete and unsuitable for commercialization. The failure of many early research prototypes of NLP based text retrieval systems has led to much skepticism in the industry, leading many to favor statistical approaches.

There has been a growing interest in the research community in the combination of NLP and conventional text retrieval. This is evidenced by the growing number of workshops on the subject. The American Association of Artificial Intelligence sponsored two of them. The first was held at the 1990 Spring AI Symposium at Stanford University on the subject of "Text Based Intelligent Systems". The second one (chaired by the applicant herein) was held at AAAI-91 in Anaheim in July 1991.

## Natural Language Techniques

The literature is rich in theoretical discussions of systems intended to provide functions similar to those outlined above. A common approach in many textbooks on natural language processing (e.g., Natural Language Understanding, James Allen, Benjamin Cummings, 1987) is to use "semantic interpretation rules" to identify the meanings of words in text. Such systems are "hand-crafted", meaning that new rules must be written for each new use. These rules cannot be found in any published dictionary or reference source. This approach is rarely employed in text retrieval is usually fails in some critical way to provide adequate results.

Krovetz has reported in various workshops (AAAI-90 Spring AI Symposium at Stanford University) and in Lexical Acquisition by Uri Zernick, Lawrence Erlbaum, 1991, ISBN 0-8056-0829-9, that "disambiguating word senses from a dictionary" would improve the performance of text retrieval systems, claiming experiments have proven that this method will improve precision. This author's philosophy suggests that a word sense be identified by "confirmation in context from multiple sources of evidence". None of Krovetz's published works propose a specific technique for doing so, and his recent publications indicate that he is "experimenting" to find suitable methods.

Eugene Charniak, of Brown University has reported in "AI Magazine" (AAAI, Winter 1992), and has spoken at the Naval Research Laboratory AI Laboratory (November 1991)about the technique of employing "spreading activation" to identify the meaning of a word in a small text. Charniak employs a "semantic network" and begins with all instances of a given word. It then "fans out" in the network to find neighboring terms that are located near the candidate term in the text. This technique suffers from 2 admitted drawbacks: it requires a high-quality partially hand-crafted, *small* semantic network, and this semantic network is *not* derived from published sources. Consequently, the Charniak method has never been applied to any text longer than a few sentences in a highly restricted domain of language.

Stephanie Haas, of the University of North Carolina, has attempted to use multiple dictionaries in information retrieval including a main English dictionary coupled with a vertical application dictionary (such as a dictionary of computer terms used in a computer database). Haas' approach does not take advantage of word sense disambiguation, and she reported at ASIS, October 1991 that merging two dictionaries gave no measurable increase in precision and recall over a single generic English dictionary.

Uri Zernick, editor of Lexical Acquisition, Lawrence Erlbaum, 1991, suggests in the same book a "cluster signature" method from pattern recognition be used to identify word senses in text. The method lists words commonly co-occurring with a word in question and determines the percentage of the time that each of the commonly occurring words appears in context in the database or corpus for each word meaning. This is called the "signature" of each word meaning. The signatures of each meaning are compared with the use of a word in context to identify the meaning. This pattern recognition approach based upon a cluster technique discussed in Duda and Hart, Pattern Classification and Scene Analysis, John Wiley & Sons, New York 1973 has the obvious drawback that it has to be "trained" for each database. The signature information is not readily obtainable from a published dictionary.

Brian Slator, (in the same book edited by Zernick above), discusses use of a "subject hierarchy" to compute a "context score" to disambiguate word senses. Generally, a "subject" or topic is identified by the context. A meaning is then selected by its relevance to the topic. This approach is only as strong as the depth of the subject hierarchy and it does not handle exceptions. A drawback of this approach is that available subject hierarchies do not cover a significant portion of the lexicon of any dictionary, let alone the vocabulary of a native speaker of a language.

One well known example of prior art in text retrieval that uses natural language input is the statistical techniques developed by Gerard Salton of Cornell University. His research system called SMART is now used in commercial applications, for example, Individual Inc. of Cambridge, MA uses it in a news clipping service. Dr. Salton is well known for his claims that natural language processing based text retrieval systems do not work as well as SMART. He bases such claims on limited experiments that he ran in the 1960's. At the 1991 ASIS meeting he stated that the reason natural language processing based systems don't work is that syntax is required and syntax is not useful without semantics. He further claims that "semantics is not available" due to the need to handcraft the rules. However, the system of the present invention has made semantics available through the use of statistical processing on machine readable dictionaries and automatic acquisition of semantic networks.

**Lexical Acquisition**

In the field of lexical acquisition, most of the prior art is succinctly summarized in the First Lexical Acquisition Workshop Proceedings, August 1989, Detroit at IJCAI-89. There is a predominance of papers covering the automatic building of natural language processing lexicons for rule-based processing. Over 30 papers were presented on various ideas, isolated concepts or prototypes for acquiring information from electronic dictionaries for use in natural language processing. None of these proposed the automatic building of a semantic network from published dictionaries.

**Indexing**

Typical text search systems contain an index of words with references to the database For a large document databases, the number of references for any single term varies widely. Many terms may have only one reference, while other terms may have from 100,000 to 1 million references. The prior art substitutes thesaurus entries for search terms, or simply requires the user rephrase his queries in order to "tease information out of the database". The prior art has many limitations. In the prior art, processing is at the level of *words,* not *concepts.* Therefore, the query explosion produces too many irrelevant variations to be useful in most circumstances. In most prior art systems, the user is required to restate queries to maximize recall. This limits such systems to use by "expert" users. In prior art systems, many relationships not found in a classical the-

saurus cannot be exploited (for example, a "keyboard" is *related to* a "computer" but it is not a synonym).

**Contextual Systems**

The prior art of systems which attempt to extract contextual understanding from natural language statements is primarily that of Gerard Salton (described in Automatic Text Processing, Addison-Wesley Publishing Company, 1989.) As described therein, such systems simply count terms (words) and co-occurrences of terms, but do not "understand" word meanings.

Routing means managing the flow of text or message streams and selecting only text that meets the desired profile of a given user to send to that user. Routing is useful for electronic mail, news wire text, and intelligent message handling. It is usually the case that a text retrieval system designed for retrieval from archived data is not good for routing and visa versa. For news wire distribution applications (which seek to automate distribution of the elements of a "live" news feed to members of a subscriber audience based on "interest profiles"), it is time-intensive and very difficult to write the compound Boolean profiles upon which such systems depend. Furthermore, these systems engage in unnecessary and repetitive processing as each interest profile and article are processed.

**Document Ranking**

Systems which seek to rank retrieved documents according to some criterion or group of criteria are discussed by Salton, in Automatic Text Processing (ranking on probabilistic terms), and by Donna Harmon, in a recent ASIS Journal article, (ranking on a combination of frequency related methods). Several commercial systems use ranking but their proprietors have never disclosed the algorithms used. Fulcrum uses (among other factors) document position and frequency. Personal Library Software uses inverse document frequency, term frequency and collocation statistics. Verity uses "accrued evidence based on the presence of terms defined in search topics".

**Concept Definition and Search**

The prior art comprises of two distinct methods for searching for "concepts". The first and most common of these is to use a private thesaurus where a user simply defines terms in a set that are believed to be related. Searching for any one of these terms will physically also search for and find the others. The literature is replete with research papers on uses of a thesaurus. Verity, in its Topic software, uses a second approach. In this approach users create a "topic" by linking terms together and declaring a numerical strength for each link, similar to the construction of a "neural network". Searching in this system retrieves any document that contains sufficient (as defined by the system) "evidence" (the presence of terms that are linked to the topic under search). Neither of these approaches is based upon the meanings of the words as defined by a publisher's dictionary.

Other prior art consists of two research programs:

- TIPSTER: A government research program called TIPSTER is exploring new text retrieval methods. This work will not be completed until 1996 and there are no definitive results to date.
- CLARIT: Carnegie Mellon University (CMU) has an incomplete prototype called CLARIT that uses dictionaries for syntactic parsing information. The main claim of CLARIT is that it indexes phrases that it finds by syntactic parsing. Because CLARIT has no significant semantic processing, it can only be viewed as a search extension of keywords into phrases. Their processing is subsumed by the present invention, with the conceptual processing and semantic networks.

**Hypertext**

Prior art electronically-retrieved documents use "hypertext", a form of manually pre-established cross-reference. The cross-reference links are normally established by the document author or editor, and are static for a given document. When the linked terms are highlighted or selected by a user, the cross-reference links are used to find and display related text.

**Machine Abstracting**

Electronic Data Systems (EDS) reported machine abstracting using keyword search to extract the key sentences based on commonly occurring terms which are infrequent in the database. This was presented at an

American Society for Information Systems (ASIS) 1991 workshop on natural language processing. They further use natural language parsing to eliminate subordinate clauses.

The present invention is similar, except that the retrieval of information for the abstract is based upon concepts, not just keywords. In addition, the present invention uses semantic networks to further abstract these concepts to gain some general idea of the intent of the document.

## Summary

The prior art may be summarized by the shortcomings of prior art systems for textual document search and retrieval. Most commercial systems of the prior art rely on "brute force indexing" and word or wild card search which provides fast response only for lists of documents which are ranked according to a precomputed index (such as document date) and not for relevance-ranked lists. For systems which attempt to relevance rank, the user must wait for the entire search to complete before any information is produced. Alternatively, some systems display documents quickly, but without any guarantee that documents displayed are the most relevant.

The systems of the prior art rank documents retrieved on the presence of words, not word meanings. The prior art systems fail to use linguistic evidence such as syntax or semantic distance. No known prior art system can combine more than a two or three ranking criteria No known system in the prior art is capable of acquiring semantic network information directly from published dictionaries, and thus, to the extent that such networks are used at all, they must be "hand built" at great expense, and with the brittleness which results from the author's purpose and bias.

In thesaurus-based information retrieval systems, as well as topic based information retrieval systems, concepts are created by linking words, not word meanings. In these systems (thesaurus and topic based), the user has the burden of creating concepts before searching. In addition, for topic based systems, the user has the added burden of making arbitrary numeric assignments to topic definitions. Prior art thesaurus and topic based systems do not link new concepts to an entire network of concepts in the natural language of search. Instead, isolated term groups are created that do not connect to the remainder of any concept knowledge base. Topic based systems require that topics be predefined to make use of concept-based processing.

Finally, for hypertext systems, authors need not spend time coding hypertext links to present a hypertextual document to users because a natural language search (perhaps taken directly from the document itself) will find all relevant concepts, not just those found by the author.

## Brief Description of the Invention

The method of the present invention combines concept searching, document ranking, high speed and efficiency, browsing capabilities, "intelligent" hypertext, document routing, and summarization (machine abstracting) in an easy-to-use implementation.

The method offers three query options:

Natural Language: finding documents with concepts expressed in plain English;
Query by Example: Present a document, retrieve similar documents;
Private Concept: define a new term, enter it in the "semantic network", search.

The method of the present invention continues to provide Boolean and statistical query options so that users will have easy access to a familiar interface and functionality while learning new and more powerful features of the present invention.

The method of the present invention is based upon "concept indexing" (an index of "word senses" rather than just words.) A word sense is a specific use or meaning of a word or idiom. The method of the present invention builds its concept index from a "semantic network" of word relationships with word definitions drawn from one or more standard English dictionaries. During query, users may select the meaning of a word from the dictionary during query construction. This results in a measurable improvement in precision.

Results of text searching are retrieved and displayed in ranked order. The ranking process is more sophisticated than prior art systems providing ranking because it takes linguistics and concepts, as well as statistics into account.

The method of the present invention uses an artificial intelligence "hill climbing" search to retrieve and display the best documents while the remainder of the search is still being processed. The method of the present invention achieves major speed advantages for interactive users.

Other significant functions of the method of the present invention including browsing documents (viewing documents directly and moving around within and between documents by related concepts), implementing "dynamically compiled" hypertext, routing, and machine abstracting or automatic summarization of long texts.

DOCKET ALARM

# Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

## Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

## Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

## Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

## API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

**LAW FIRMS**
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

**FINANCIAL INSTITUTIONS**
Litigation and bankruptcy checks for companies and debtors.

**E-DISCOVERY AND LEGAL VENDORS**
Sync your system to PACER to automate legal marketing.