PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 17/30 | A1

(11) International Publication Number: **WO 98/16890**

(43) International Publication Date: 23 April 1998 (23.04.98)

(21) International Application Number: PCT/US97/18712

(22) International Filing Date: 14 October 1997 (14.10.97)

(30) Priority Data:
60/028,437 15 October 1996 (15.10.96) US

(71) Applicant *(for all designated States except US)*: MANNING & NAPIER INFORMATION SERVICES [US/US]; 1100 Chase Square, Rochester, NY 14604 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: SNYDER, David, L. [US/US]; 52 Crestview Drive, Pittsford, NY 14534 (US). CALISTRI-YEH, Randall, J. [US/US]; 495 Pellett Road, Webster, NY 14580 (US).

(74) Agent: DURDIK, Paul, A.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111-3834 (US).

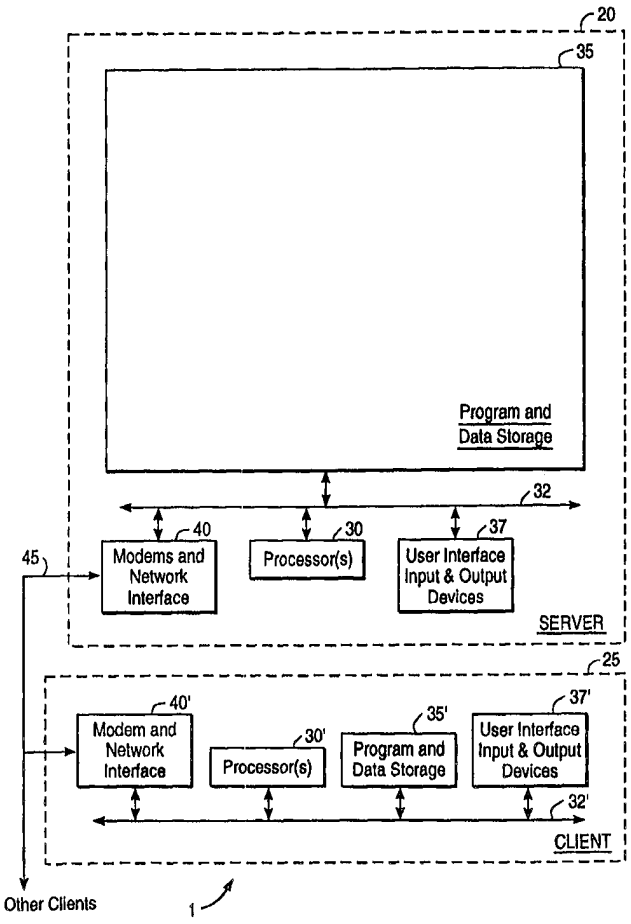
(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: MANAGEMENT AND ANALYSIS OF DOCUMENT INFORMATION TEXT

(57) Abstract

An interactive system for analyzing and displaying information (Fig. 1A) contained in a plurality of documents employing both term-based analysis and conceptual-representation analysis (Fig. 9D). Particulars of the invention are especially effective for analyzing patent texts, such as patent claims, abstracts and other portions of a patent document.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

MANAGEMENT AND ANALYSIS OF DOCUMENT INFORMATION TEXT

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority from the following U.S. Provisional Application:

U.S. Provisional Patent Application, serial no. 60/028,437, David L. Snyder and Randall J. Calistri-Yeh, entitled, "Management and Analysis of Patent Information Text (MAPIT)", filed October 15, 1996.

CROSS-REFERENCE TO ARTICLES

The following publications are directed to techniques for measuring document similarity including information directed to subject field coders, semantic thread analysis and/or TF.IDF techniques:

Liddy, E.D., Paik, W., Yu, E.S. & McVearry, K., "An overview of DR-LINK and its approach to document filtering," Proceedings of the ARPA Workshop on Human Language Technology (1993);

Liddy, E.D. & Myaeng, S.H. (1994). DR-LINK System: Phase I Summary. Proceedings of the TIPSTER Phase I Final Report.

Liddy, E.D., Paik, W., Yu, E.S. & McKenna, M. (1994). Document retrieval using linguistic knowledge. Proceedings of RIAO '94 Conference.

Liddy, E.D., Paik, W., Yu, E.S. Text categorization for multiple users based on semantic information from an MRD. ACM Transactions on Information Systems. Publication date: 1994. Presentation date: July, 1994.

Liddy, E.D., Paik, W., McKenna, M. & Yu, E.S. (1995) A natural language text retrieval system with relevance feedback. Proceedings of the 16th National Online Meeting.

Paik, W., Liddy, E.D., Yu, E.S. & McKenna, M. Categorizing and standardizing proper nouns for efficient information retrieval. Proceedings of the ACL Workshop on Acquisition of Lexical Knowledge from Text. Publication date: 1993.

Paik, W., Liddy, E.D., Yu, E.S. & McKenna, M. Interpretation of Proper Nouns for Information Retrieval. Proceedings of the ARPA Workshop on Human Language Technology. Publication date: 1993.

Salton, G. and Buckley, C. Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management. Volume 24, 513-523. Publication date: 1988 ("Salton reference").

BACKGROUND OF THE INVENTION

The present invention relates to information management and, more particularly to the management and analysis of document information

text.

We live in the information age. How prophetic the statement of a major computer manufacturer which said "It was supposed to be the atomic age, instead it has turned out to be the information age." Prophetic both in the impact of the age, as well as its potential for beneficial and deleterious effects on humankind. Faced with an explosion of information fueled by the burgeoning technologies of networking, inter-networking, computing and the trends of globalization and decentralization of power, today's business manager, technical professional and investment manager are faced with the need for careful, accurate and timely analysis of the deluge of information underlying their everyday decisions. Several factors underlie this need for prompt information analysis. First, in an era of ever tighter cost controls and budgetary constraints, companies are faced with a need to increase their operational efficiency. In doing so, they face the need to assimilate large amounts of accounting and financial information, both concerning their internal functioning as well as their position in the market place. Second, the omnipresent factor of litigation which may cost or earn a company billions of dollars. The outcome of such contests is often determined by which side has access to the most accurate information. Third, the drive for greater economies of scale and cost efficiencies spurs mergers and acquisitions, especially in high technology areas. The success of such activity is highly dependent upon who has superior abilities to assimilate information. Fourth, the explosive growth of technology in all areas, especially in biotechnology, computing and finance, brings with it the need to access and comprehend technical trends impacting the individual firm. Fifth, the globalization of the market place in which today's business entities find themselves brings with it the need to master information concerning a multiplicity of market mechanisms in a multiplicity of native languages and legal systems. Sixth, the decentralization of large industrial giants has led to the need for greater cross-licensing of indigenous technologies; requiring that companies discern precisely the quantity and kinds of technology being cross-licensed.

Faced with the increasing importance of successful analysis of a burgeoning information stockpile, today's business professional is faced, as never before, with a need for tools which not only find information, but find the correct information, as well as, assist the user in drawing conclusions and perceiving the meaning behind the information resources discovered.

The most typical information analysis tool available today is a database of text or images which is searched by a rudimentary search engine. The user enters a search query consisting of specific key words encoded in a boolean formalism. Often the notation is so complex that trained librarians are needed to ensure that the formula is correct. The results of database searches are a list of documents containing the key words the user has requested. The user often does not know the closeness

of the match until each reference cited by the search engine is studied manually. There is often no way to search different portions of documents. Finally, the output of this process is a flat amalgam of documents which has not been analyzed or understood by the system performing the search.

The user who turns to an automated information analysis system is seeking not merely a collection of related documents, but the answers to critical questions. For example,

"Are there any issued patents that are so close to this invention proposal that they might represent a potential infringement problem?"

"Are the resources of company X complimentary to our own company such that we should consider a merger with company X?"

"Of the court cases decided in California last year, how many of them involved a sexual harassment charge?"

"What companies exist as potential competitors in the market place for our planned product?"

Current analysis tools demonstrate themselves to be ineffective when faced with these types of issues. What is needed is an information analysis tool capable of analyzing, acquiring, comprehending a large amount of information and presenting that information to users in a intelligible way.

SUMMARY OF THE INVENTION

The present invention provides an interactive document-management-and-analysis system and method for analyzing and displaying information contained in a plurality of documents. Particular embodiments of the invention are especially effective for analyzing patent texts, such as patent claims, abstracts, and other portions of the specification.

A method according to one embodiment of the invention includes generating a set of N different representations of each document, and for each of a number of selected pairs of documents, determining N utility measures, a given utility measure being based on one of the N representations of the documents in that pair. In a specific embodiment, this information is displayed as a scatter plot in an area bounded by N non-parallel axes, where each selected pair is represented by a point in N-space having its coordinates along the N axes equal to the N utility measures.

In a specific embodiment, wherein N=2, the first representation is a conceptual-level representation such as a subject vector, and the second representation is a term-based representation such as a word vector.

In one use scenario, the selected pairs include all pair wise combinations of the documents in the plurality. In another scenario, the selected pairs are all pair wise combinations that include a particular document in the plurality.

The use of multiple methods of analysis, such as, for example, word-vector analysis and semantic-thread analysis, creates synergistic benefits by providing multiple independent measures of similarity. A system which uses multiple methods together can discover similar documents that either single method may have overlooked. In the cases where both methods agree, the user has greater confidence in the results because of the built-in "second opinion".

In accordance with another aspect of the invention, a dynamic concept query is performed by treating a user-specified query as a special type of document. The user can enter a list of words ranging from a single keyword to the text of an entire document, which is treated as a new document. A multidimensional array of similarity scores comparing that document to each existing document in the set is calculated. The user can then view the resulting clusters using the visualization techniques described herein.

The invention provides for an innovative analysis tool that assists users in discovering relationships among thousands of documents such as patents. Sophisticated natural language and information retrieval techniques enable the user to analyze claim sets, cluster claims based on similarity, and navigate through the results using graphical and textual visualization.

The invention provides further for a search routine which goes beyond simple keyword search; it understands the structure of documents such as patents and it captures concepts like patent infringement and interference. Users can browse through data visualizations (e.g., range query as described below), inspect quantitative score comparisons, and perform side-by-side textual analysis of matching patent claims. Based on the information gathered, users may analyze competitive patent and acquisition portfolios, develop patent blocking strategies, and find potential patent infringement.

In accordance with another aspect of the invention, the analysis methods described herein may be applied to a set of documents formed by the additional step of filtering a larger set of documents based on a concept query. For example, a user may find it useful to examine only those patents that discuss microelectronic packaging, analyze those patents, and generate a scatter plot display (e.g., run a concept query to pick a claim of interest followed by a claim query based on such claim and generate an overlay plot as described below).

In accordance with another aspect of the invention, recognizing and exploiting the relationship between various document types and "compound documents" each to the other permits multi-faceted analyses of multiple document types. For example, a patent is a compound document with nested sub-document linkages to sub-components, such as claims, background and summary of invention, etc. A claim is also a compound document because it may refer to other claims. Applying this paradigm to document analysis strategies, claims, whether individual, nested or as an

amalgam, may be compared to other compound document components. For example, comparing claims to background and summary of invention in a patent. Furthermore, claims and background and summary of invention can be compared to other documents, such as related prior art literature from other sources such as magazines or journals. This enables the patent practitioner to view relevant claims, background and summaries, and other documents (non-patents), and cluster these together by similarity measures.

In accordance with one aspect of the invention, the user may select a metric that captures the essence of the document or documents under analysis. For example, the legal concept of patent infringement may be applied to sets of patents or patent applications. In a particular embodiment, a similarity matching algorithm treats the exemplar part of a patent claim differently from the dependent parts of the claim. Thus, a kind of "cross-comparison" matching is used, wherein the combined scores for (1) patent A, claim X dependent and independent part(s) vs. patent B, claim Y, independent part and (2) patent A, claim X dependent and independent part(s) vs. patent B, claim Y, dependent and independent part(s), generate an aggregate matching (or similarity) score for patent A, claim X vs. patent B, claim Y.

Normalization techniques deal with asymmetries in the matching, especially for documents of different lengths. For example, in the patent context, the situation where there is a short claim on "blue paint" and a long claim containing "blue paint." Looking at the small claim vs. the long claim appears close (since the long one at least contains the small one). But what of the case where it's the long claim vs. the small one? Standard information retrieval techniques would dictate that it's a poor match, since the long claim contains many limitations not in the small one. For patents, the "interference/infringement" match suggests that these are close, because if one "covers" the other, it doesn't matter which is the "query" and which is the "document."

Similarity based on the legal concept of patent infringement and interference serves as the touchstone to analyze, cluster and visualize patents and applications. This enables users to evaluate incoming applications for infringement against existing patents, filter large sets of patents to remove reissued and derivative patents, identify significant claim modifications in a reissued patent and identify related and unrelated patents to compare the intellectual property of two businesses.

A further understanding of the nature and advantages of the present invention may be realized by reference to the remaining portions of the specification and the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1A is a block diagram of a document analysis system

embodying the present invention;

Fig. 1B is a more detailed block diagram of the interactions between the user and the system during the processing of document information;

Figs. 2A-2B depict off line structured document processing steps according to a particular embodiment;

Fig. 3 depicts the preprocess step of off line structured document processing of Figs. 2A-2B according to a particular embodiment;

Fig. 4A depicts the mapit-process step of off line structured document processing of Figs. 2A-2B according to a particular embodiment;

Fig. 4B depicts the mapit-sfc step of Fig. 4A according to a particular embodiment;

Fig. 5 depicts the on line concept query processing according to a particular embodiment;

Figs. 6A-6B depict off line generic document processing steps according to a particular embodiment;

Fig. 7A depicts claim parsing according to a particular embodiment;

Fig. 7B depicts the process-words step of claim parsing of Fig. 7A according to a particular embodiment;

Fig. 8A illustrates a scatter plot visualization technique according to a particular embodiment of the invention;

Fig. 8B illustrates a 2D plot visualization technique according to a particular embodiment of the invention;

Fig. 8C illustrates a 3D plot visualization technique according to a particular embodiment of the invention;

Fig. 8D illustrates an S-curve plot visualization technique according to a particular embodiment of the invention;

Fig. 8E is a flow chart depicting the steps for generating an S-curve plot;

Fig. 9A illustrates a representative sign on screen according to a particular embodiment of the invention;

Fig. 9B illustrates a representative dataset select screen according to a particular embodiment of the invention;

Fig. 9C illustrates a representative concept query screen according to a particular embodiment of the invention;

Fig. 9D illustrates a representative concept query review screen according to a particular embodiment of the invention;

Figs. 9E and 9F illustrate representative concept query results screens according to a particular embodiment of the invention;

Fig. 9G illustrates a representative concept query results viewer screen according to a particular embodiment of the invention;

Figs. 9H and 9I illustrate representative concept query results viewer screens depicting side-by-side comparisons according to a particular embodiment of the invention;

Fig. 9J illustrates a representative claim viewer screen

according to a particular embodiment of the invention;

Fig. 9K illustrates a representative patent viewer screen according to a particular embodiment of the invention;

Fig. 10A illustrates a representative patent query screen according to a particular embodiment of the invention;

Fig. 10B illustrates a representative patent query results screen according to a particular embodiment of the invention;

Fig. 10C illustrates a representative patent query side-by-side comparison screen of claims according to a particular embodiment of the invention;

Fig. 10D illustrates a representative patent query side-by-side comparison screen of patents according to a particular embodiment of the invention;

Fig. 11A illustrates a representative claim query screen according to a particular embodiment of the invention;

Fig. 11B illustrates a representative claim query claim finding screen according to a particular embodiment of the invention;

Fig. 11C illustrates a representative claim query results screen according to a particular embodiment of the invention;

Fig. 11D illustrates a representative claim query side-by-side comparison screen of claims according to a particular embodiment of the invention;

Fig. 11E illustrates a representative overlay plot for a claim query results screen according to a particular embodiment of the invention;

Fig. 11F is a flow chart depicting the steps for generating an overlay plot;

Fig. 12A illustrates a representative range query screen according to a particular embodiment of the invention;

Fig. 12B illustrates a representative range query results screen according to a particular embodiment of the invention;

Fig. 12C is a flow chart depicting the steps for generating a range query; and

Figs. 13A-13G illustrate an alternative embodiment of the present invention.

DESCRIPTION OF SPECIFIC EMBODIMENTS

A preferable embodiment of a document-management-and-analysis system and method according to the invention applicable to the task of patent search and analysis is reduced to practice and is available under the trade name, MAPIT™.

A document search and analysis tool must be both fast enough to handle a voluminous quantity of documents and flexible enough to adapt to different user requirements. Other aspects of the invention are of particular importance to expedient, accurate and efficient document analysis. First, the understanding of the structure and content of

documents on multiple levels is useful to provide a much deeper analysis than generic search engines known in the art. Second, the combination of multiple similarity metrics is useful to achieve highly customized results. By contrast, search engines known in the art restrict the user to whatever notion of similarity was incorporated into the system by its designers. Third, the ability to parse structured documents, such as patent claims, is useful to extract their meaning. Fourth, the graphical display of information relevant to the user provides the user with quick access to the product of the analysis.

In accordance with the invention, multiple forms of textual analysis used to compare documents may be combined in any particular embodiment. One textual analysis method, called word-vector (also referred to as "wordvec" and "term-based") analysis, focuses on the co-occurrences of individual words and phrases between documents under analysis. Stemming technology, used in conjunction with word-vector analysis, matches words such as "projected" and "projection". Noun phrases, which are the key building blocks for many documents, such as patents, are identified and isolated. Those technologies are not restricted to English, but can be applied directly to other European languages. Word-vector analysis may be reduced to practice using Term Frequency/Inverse Document Frequency ("TF.IDF") techniques, or other techniques known in the art. TF.IDF techniques are further described in the Salton reference identfied above.

Another form of textual analysis is called semantic thread analysis (also referred to herein as "subject-vector" or "conceptual representation" analysis). Instead of focusing on individual words, this method identifies the general topics and themes in a document. It can determine that a patent, for example, is 35% about engineering physics, 15% about polymer science, 20% about holography, and 30% about manufacturing processes. If two patents cover the same subject areas in the same proportions, it is likely that they are closely related even if they use completely different words to describe their inventions. Semantic thread analysis may be reduced to practice by employing subject field code (SFC) techniques described by Dr. Elizabeth Liddy, et al. in one or more of the references identified above.

Preface on the Format of the Drawings

Embodiments of the invention will be best understood with reference to the drawings included herewith. A note on the format of these drawings is in order. In the drawings, process steps are depicted as squares or rectangles. Data structures internal to the program are depicted as rhomboid type structures. For example, in reference to Fig. 2A, element 10, *text format file of patents in a search set* is a rhomboid structure. Conventional data or text files are depicted as squares or rectangles with the upper right hand corner turned downward. For example, in Fig. 2A element 50 *justclaims* is a file which may exist on a hard disk,

floppy disk, CD ROM or other form of storage medium. Open ended arrows reflect the flow of information. Tailess arrows indicate the flow of processing.

These drawings depict the processing steps, files and information according to one embodiment of the invention targeted to processing and understanding patents. While this serves as an excellent example of the features of the invention, the reader with ordinary skill in the art will appreciate that the invention's scope encompasses not merely the understanding and analysis of patents, but other documents as well.

Table 1 provides a definitional list of terminology used herein.

Term	Definition
Claim Query	A query against a collection of text documents compared to a part of a particular member of the collection.
Concept Query	A query against a collection of text documents compared to a user input textual concept.
Corpus	A dataset.
Dataset	A document database containing documents upon which search and analysis operations are conducted.
Document	A unit of text which is selected for analysis which may include an entire document or any portion thereof such as a title, an abstract, or one or more clauses, sentences, or paragraphs. A document will typically be a member of a document database containing a large number of documents and may be referred to by the term corpus.
DR LINK	Document Retrieval using LInguistic Knowledge. This is a system for performing natural language processing. This system is described in papers by Dr. Liddy referenced in the cross-reference section herein above.
Patent Query	A query against a collection of text documents compared to a particular member of the collection, identified by the user.
Polysemy	The ability of a word to have multiple meanings.
Query	Text that is input for the purpose of selecting a subset of documents from a document database. While most queries entered by a user tend to be short compared to most documents stored in a database this should not be assumed.
score	A numerical indicator assigned to a document indicative of a particular characteristic, e.g. relevance to a query.
Searchset	A document database containing documents upon which search and analysis operations are conducted.
SFC	Subject field coder. A subject field coder is a process which tags content-bearing words in a text with a disambiguated subject code using a lexical resource of words which are grouped in subject categories.

SGML	Standard Generalized Markup Language. Standard generalized markup language is comprised of a set of tags which may be embedded into a text document to indicate to a text processor how to process the surrounding or encompassed text.
Split Dataset	A dataset may be split into two distinct components in order to perform comparative analyses between the two sub-datasets. For example, a split dataset of A company patents and B company patents enables the user to discover relationships between the patent portfolios of the two companies.
Stemming	Stemming is a process whereby nouns are reduced to their most basic form or stem. For example, the words "processing" and "processed" are stemmed to the word "process".
Stop Word	One of a collection of words which are not assigned a semantic meaning by the system. For example, the word "the".
Stop Word	List A list of stop words.
Term Index	A unique identifier assigned to each stem by a term indexer.
Term Indexer	Term indexer is a process which performs indexing on an input text. Indexing involves extracting terms from the text, checking for stop words, processing hyphenated words, then stemming all inflected terms to a standard form. Finally, a unique term index is assigned to each stem.
TFIDF	Term Fequency/Inverse Document Frequency. This is a score computed by a term indexer process. This score determines the relative prominence of a term compared to its occurrence throughout a document body.
Token	A white space delimited sequence of characters having a particular meaning.
Tokenize	A process whereby input text is separated into a collection of tokens.
Transitive Closure	The transitive closure of a claim is the claim itself and the transitive closure of all references within the particular claim.
weight	A numerical indicator assigned to a word or token indicative of a particular characteristic, e.g. relevance to a query.
Word	A single word, compound word, phrase or multiword construct. Note that the terms "word" and "term" are used interchangeably. Terms and words include, for example, nouns, proper nouns, complex nominals, noun phrases, verbs, and verbs numeric expressions and adjectives. These include stemmed and non-stemmed forms.

TABLE 1

Hardware Overview

The document-management-and-analysis system (the "system") of the present invention is implemented in the "C", "C++", "perl" and UNIX shell script programming languages and is operational on a computer system such as shown in Fig. 1A. This figure shows a conventional client-server computer system 1 that includes a server 20 and numerous clients, one of which is shown at 25. The use of the term "server" is used in the context of the invention, where the server receives queries from (typically remote) clients, does substantially all the processing necessary to formulate responses to the queries, and provides these responses to the clients. However, server 20 may itself act in the capacity of a client when it accesses remote databases located on a database server. Furthermore, while a client-server configuration is known, the invention may be implemented as a standalone facility, in which case client 25 would be absent from the figure.

The hardware configurations are in general standard, and will be described only briefly. In accordance with known practice, server 20 includes one or more processors 30 that communicate with a number of peripheral devices via a bus subsystem 32. These peripheral devices typically include a storage subsystem 35 (memory subsystem and file storage subsystem holding computer program (e.g., code or instructions) and data implementing the document-management-and-analysis system), a set of user interface input and output devices 37, and an interface to outside networks, including the public switched telephone network. This interface is shown schematically as a "Modems and Network Interface" block 40, and is coupled to corresponding interface devices in client computers via a network connection 45.

Client 25 has the same general configuration, although typically with less storage and processing capability. Thus, while the client computer could be a terminal or a low-end personal computer, the server computer would generally need to be a high-end workstation or mainframe, such as a SUN sparc server. Corresponding elements and subsystems in the client computer are shown with corresponding, but primed, reference numerals.

The user interface input devices typically includes a keyboard and may further include a pointing device and a scanner. The pointing device may be an indirect pointing device such as a mouse, trackball, touchpad, or graphics tablet, or a direct pointing device such as a touchscreen incorporated into the display. Other types of user interface input devices, such as voice recognition systems, are also possible.

The user interface output devices typically include a printer and a display subsystem, which includes a display controller and a display device coupled to the controller. The display device may be a cathode ray tube (CRT), a flat-panel device such as a liquid crystal display (LCD), or a projection device. Display controller provides control signals to the display device and normally includes a display memory for storing the

pixels that appear on the display device. The display subsystem may also provide non-visual display such as audio output.

The memory subsystem typically includes a number of memories including a main random access memory (RAM) for storage of instructions and data during program execution and a read only memory (ROM) in which fixed instructions are stored. In the case of Macintosh-compatible personal computers the ROM would include portions of the operating system; in the case of IBM-compatible personal computers, this would include the BIOS (basic input/output system).

The file storage subsystem provides persistent (non-volatile) storage for program and data files, and typically includes at least one hard disk drive and at least one floppy disk drive (with associated removable media). There may also be other devices such as a CD-ROM drive and optical drives (all with their associate removable media). Additionally, the computer system may include drives of the type with removable media cartridges. The removable media cartridges may, for example be hard disk cartridges, such as those marketed by Syquest and others, and flexible disk cartridges, such as those marketed by Iomega. One or more of the drive may be located at a remote location, such as in a server on a local area network or at a site of the Internet's World Wide Web.

In this context, the term "bus subsystem" is used generically so as to include any mechanism for letting the various components and subsystems communicate with each other as intended. With the exception of the input devices and the display, the other components need not be at the same physical location. Thus, for example, portions of the file storage system could be connected via various local-area or wide-area network media, including telephone lines. Similarly, the input devices and display need not be at the same location as the processor, although it is anticipated that the present invention will most often be implemented in the context of PCs and workstations.

Bus subsystem 32 is shown schematically as a single bus, but a typical system has a number of buses such as a local bus and one or more expansion buses (e.g., ADB, SCSI, ISA, EISA, MCA, NuBus, or PCI), as well as serial and parallel ports. Network connections are usually established through a device such as a network adapter on one of these expansion buses or a modem on a serial port. The client computer may be a desktop system or a portable system.

The user interacts with the system using interface devices 37' (or devices 37 in a standalone system). For example, client queries are entered via a keyboard, communicated to client processor 30', and thence to modem or network interface 40' over bus subsystem 32'. The query is then communicated to server 20 via network connection 45. Similarly, results of the query are communicated from the server to the client via network connection 45 for output on one of devices 37' (say a display or a printer), or may be stored on storage subsystem 35'.

Fig. 1B is a functional diagram of computer system 1. Fig. 1B depicts a server 20 preferably running Sun Solaris software or its equivalent, and a representative client 25 of a multiplicity of clients which may interact with the server 20 via the internet 45 or any other communications method. Blocks to the right of the server are indicative of the processing steps and functions which occur in the server's program and data storage indicated by block 35 in Fig. 1A. Input search set 10 which in this embodiment is a text format file of patents to be searched serves as the input to query processing block 35A. Query processing manipulates the input data 10 to yield a searchable dataset 10A. A Common Gateway Interface (CGI) script 35B enables queries from user clients to operate upon the dataset 10A and responses to those queries from the information in the dataset 10A back to the clients in the form of a Hypertext Markup Language (HTML) document outputs which are then communicated via internet 45 back to the user.

Client 25 in Fig. 1B possesses software implementing the function of a web browser 35A´ and an operating system 35B´. The user of the client may interact via the web browser 35A´ with the system to make queries of the server 20 via internet 45 and to view responses from the server 20 via internet 45 on the web browser 35A´.

In accordance with one aspect of the invention, documents may be thought of as belonging to two broad categories. The first are structured documents; those having a very highly specific structure. For example the claims incorporated within patents are structured. To accurately compare claims from two different patents, it is necessary to realize that a claim may refer to earlier claims, and those earlier claims must enter into the analysis. Furthermore, for purposes of infringement analysis, it is important to treat the "head" of a chain of dependent claims differently from the rest of the body. The second type of document is a more generic form of document having no definable structural components referred to as generic documents.

In a particular embodiment, the invention divides overall processing into an off-line processing step and an on-line query step. The off-line processing step will process incoming document information from a variety of input sources, such as a database of U.S. patents, a collection of documents scanned into electronic format by a scanner or a database of newsprint, and build from it structures which allow the system to be able to manipulate and interpret the data acquired from these input sources. The query steps on the other hand, are targeted to on-line interactions with the system to gain from it knowledge about information which the off-line step has processed.

Off-line processing

Fig. 2A depicts off-line processing of structured documents (or claims in this example) in this particular embodiment of the invention. A text format file of patents search set 10 comprises the

input data to the system. Input data may be in multiple formats.

Claim processing for a data set begins with the step of creating a justclaims file 50 for each patent in the set, pursuant to step 102 of Fig. 2A. Each file 50 contains the text of all the claims of one patent disposed within the set. The reader of ordinary skill in the art will appreciate that the specific processing of this step necessarily conforms to the format of the source text available to the system. For example, if the source text is in text format, this step must process textual data. Next a justclaimslist 52 is produced in step 104. The justclaimslist contains the full directory path to each justclaims file 50 in the order that they are processed.

Pursuant to step 106, a make-claims routine is executed. This make-claims routine takes each of the justclaims files 50 created in step 102 and one line from justclaimslist 52 and creates two separate files for each claim contained in file 50 (and therefore in a patent). The first file, called single file 54, contains the text of one claim. The second file, called merged file 56, contains the text of one claim plus the text of the transitive closure of all claims referenced by that claim. The output from make-claims step 106 also includes a claimlist data structure 12 and a patentlist data structure 14 (in the form of conventional binary data file). The make-claims step employs numerous heuristics in an attempt to identify both the scope and references of the claims. For example: 1) Each claim must start on a new line and that line must start with the claim number, a period and one or more spaces; 2) Claims must be numbered sequentially starting with 1 (note that this heuristic will not catch the case where, for example, claim 4 has text including a line starting with "5."); 3) References to other claims are understood by the system, such as: a) "claim 3", b) "Claim 3", c) "claim 2 or 3", d) "claim 2 and 3", e) "claims 2 or 3", f) "claims 2 and 3", g) "claims 2, 3, or 4", h) "claims 2-4", i) "claims 2 to 4", j) "claims 2 through 4", k) "claims 2-5 inclusive or 8", l) "all previous claims", m) "any proceeding claims"; and 4) Claims can only refer to claims occurring previously in documents. It is possible but rare to legally refer to a future claim. It is rather common to have a typographic error refer to a future claim by mistake. If a reference to a future claim is encountered, a warning message is printed, the reference is skipped and processing continues. (This warning is forwarded to the user, who determines whether the reference to a future claim is intentional or a typographical error.) All claims referred to by the current claim, and all claims recursively referred to by any of them, are printed in the order encountered following the text of the current claim. The remaining heuristics are specified in Table 3 below.

Preprocess step 108 has a task of taking raw document input, filtering from it extraneous matter and extracting root words and noun phrases. A commercial-off-the- shelf (COTS) language processing tool such as the XLT software package available from Xerox, a corporation with

headquarters in Stamford, Connecticut, performs much of the processing. Although other software such as part-of-speech (POS) tagger of the type provided by such companies as Inso Corporation, Boston, Massachusetts may also be used. Its behavior in this embodiment is depicted with greater particularity in Fig. 3.

Referring to flowchart 201 of Fig. 3, the preprocess step initially prefilters input text and removes nonlegible items, pursuant to step 200. Any number of appropriate heuristics may be used, such as dropping any words with more than fifty characters.

Next, a tokenize step 210 tokenizes the document text. Following step 210, all words are converted to lower case pursuant to step 220. Each word is then reduced to its derivational root in stem step 230. For example, the words "processed" and "processing" would be stemmed to the word "process". Stems are written out in step 240 with two exceptions: 1) If the stemmed word contains anything except letters, it is not printed and 2) If the original word is contained in the stop word list, it is not printed.

The next step is tag words step 250. All words are tagged with their part of speech working one sentence at a time. If the sentence has more than 1,000 tokens, the program will skip this sentence. Post filter step 260 removes phrases suspected of being in error. For example, known phrases with more than five nouns in a row are removed. I.D. noun phrases 270 removes extraneous noun-phrases. For example, if the phrase contains the word "said" the phrase is removed. If the phrase contains the word "claim" or "claims", the phrase is removed. Additional extraneous terms related specifically to the subject technology may also be identified and removed.

In step Write-out noun phrases 280, all noun phrases are written to the standard output on a single line separated by a space. The words in the phrase are joined by an underscore ("_"). In summary, preprocess step 108 produces a single file for each input document containing the foregoing subject matter ("preprocessing text file"), representing preprocessed documents for subsequent analysis.

Referring again to Fig. 2A, after preprocess step 108 completes, processing continues with a build-claimlist step 110. Build-claimlist translates the full directory paths of the justclaims files 50 represented in the justclaimslist file 52 as ASCII directory paths into a binary represented form of the directory path information stored in the claimlist.bin file 49. This enables later processing to work with binary represented full directory paths for these files, which is more efficient than working with the text represented files.

Following step 110, a build-hash step 112 creates a series of hash files 60 that enable the system to rapidly access information about various documents and claims. Each hash file consists of two separate files containing mapping information linking together information about the documents being processed. These hash files 60 are: 1) A mapping

from a claim number to a unique document index, representative of each document being analyzed; 2) A mapping from a claim number to the full directory path to the text of that claim; 3) A mapping from a claim number to the first 150 characters of the claim; 4) A mapping from a patent number to the unique document index; 5) A mapping from a patent number to a full directory path to the text of that patent; 6) A mapping from a patent number to the full title of the patent; 7) A mapping from a patent number to the assignee of that patent; and 8) A mapping from a patent number to a space separated list of claims included in that patent. Each hash file created in step 112 is a mapping between an ASCII key string and an ASCII value string.

Following step 112, a fix-patentlist step 114 removes entries from patentlist data structure 14 that do not have any claims. The original patentlist is backed up to an original patentlist file 16. The remaining good patents (i.e., those with claims) stay in patentlist structure 14. Any bad patents are written to a separate data structure 18. Processing now continues with a mapit-process step 116 which is described in greater detail in flowchart 301 of Fig. 4A.

Referring to flowchart 301, initially tf step 300 translates a set of claimlist text files 12 which have been processed by the preprocess step 108 in Fig. 2A into a single file 64, which consists of a list of each unique term in the original claimlist files followed by a count of the number of occurrences of that term for each document. This file is the last ASCII represented file that is produced during processing.

Step 310 next takes the file 64 produced by step 300 and creates four files 66 used in calculations in tfidf-all step 320. Included in files 66 are: 1) A hash file mapping each term in the body of documents being analyzed to a unique index; 2) A binary file containing a single integer value for the number of words in the hash file; 3) A binary version of the file created by step 300 recording the term index and the frequency count for each term in each document; and 4) A mapping of an unique index associated with each term to the number of documents that contain that term. The total number of terms including duplicates in each document is printed to a standard-out (STDOUT) and is typically redirected a to convenient file.

Referring again to Fig. 4A, step 320 calculates actual TFIDF weights for each term in each document in the claimlist producing a file of weights 72. These weights are combined by mapit-all step 120 (Fig. 2B) or mapit-process-query step 420 (Fig. 5) to generate a "score" for a pair of documents. In a preferable embodiment two separate sets of weights are calculated for each document. The first set, query weights, is to be used when comparing the document against a concept query. The second set, doc weights, is used when comparing a document against another document. TF.IDF technqiues for calculating doc weights are further described in Salton reference identified above.

Following tfidf_all step 320, a normalize step 330 calculates

a set of normalization factors that force all document-pair scores to lie between 0.0 and 1.0. By definition a document compared against itself as a perfect score of 1.0 and no other document can score higher than 1.0. In a preferable embodiment, a document is scored against itself by calculating the term weights with formula (4) hereinabove and then taking the dot product to arrive at a normalization factor. A score for this document against any other document is divided by this normalization factor, yielding a maximum score of 1.0.

After step 330, a make-twfmt step 340 creates an SFC input file 68 for processing by a Subject Field Coder ("SFC"). A Subject Field Coder (SFC) tags content-bearing words in a text with a disambiguated subject code using a lexical resource of words whose senses are grouped by subject categories.

A subject field code indicates the conceptual-level sense or meaning of a word or phrase. The present invention, however, is not limited to a specific hierarchical arrangement or a certain number or scheme of subject field codes.

Each information bearing word in a text is looked up in a lexical resource. If the word is in the lexicon, it is assigned a single, unambiguous subject code using, if necessary, a process of disambiguation. Once each content-bearing word in a text has been assigned a single SFC, the frequencies of the codes for all words in the document are combined to produce a fixed length, subject-based vector representation of the document contents. This relatively high-level, conceptual representation of documents and queries is a useful representation of texts used for later matching and ranking.

Polysemy (the ability of a word to have multiple meanings) is a significant problem in information retrieval. Since words in the English language have, on average, about 1.49 senses, with the most commonly occurring nouns having an average of 7.3 senses, and the most commonly occurring verbs having an average of 12.4 senses, a process of disambiguation is involved in assigning a single subject field code to a word.

Words with multiple meanings (and hence multiple possible subject field code assignments) are disambiguated to a single subject field code using three evidence sources (this method of disambiguation has general application in other text processing modules to help improve performance):

Local Context: If a word in the sentence has been tagged with only one concept group code, this concept group code is considered Unique. Further, if there are any concept group codes which have been assigned to more than a predetermined number of words within the sentence being processed, these concept group codes are considered Frequent codes. These two types of locally determined concept group codes are used as "anchors" in the sentence for disambiguating the remaining words. If any of the ambiguous (polysemous) words in the sentence have either a Unique

or Frequent concept group code amongst their codes, that concept group code is selected and that word is thereby disambiguated.

Domain Knowledge: Domain Knowledge representations reflect the extent to which words of one concept group tend to co-occur with words of the other concept groups (hence the notion of the domain predicting the sense). For example, within a given sentence, a word with multiple concepts categories is disambiguated to the single concept category that is most highly correlated with the Unique or Frequent concept category. If several Unique or Frequent anchor words exist, the ambiguous word is disambiguated to the correct category of the anchor word with the highest overall correlation coefficient.

Global Knowledge: Global Knowledge simulates the observation made in human sense disambiguation that more frequently used senses of words are cognitively activated in preference to less frequently used senses of words. Therefore, the words not yet disambiguated by Local Context or Domain Knowledge will now have their multiple concept group codes compared to a Global Knowledge database source.

Subject field codes are further discussed in Liddy, E.D., Paik, W., Yu, E.S. & McVearry, K., "An overview of DR-LINK and its approach to document filtering," Proceedings of the ARPA Workshop on Human Language Technology (1993).

Processing in step 340 concatenates all claim files together (e.g., single file 54 or merged file 56, etc.) and adds several Standard Generalized Mark-up Language ("SGML") tags as are well known in the art. (Such processing is described in greater detail by Dr. Liddy, et al. in "Categorizing And Standarizing Proper Nouns For Efficient Information Retrieval").

Note that since documents are represented by SFCs, which are language independent, a related embodiment can perform multi-language word vector analysis on sets of documents. Thus, a related embodiment could, for example, analyze a set of French patents.

Mapit-sfc step 350 next performs subject field coding on the SFC input file 68 produced in step 340. Processing mapit-sfc step 350 is detailed in flowchart 361 of Fig. 4B. Referring to Fig. 4B, the first step of such processing is dpfilter step 360 which removes unwanted SGML delimited text. Following step 360, sfc-tagger step 370 uses a part of speech tagger to parse all documents one sentence at a time. Sfc step 380 identifies subject field codes and the weighting for each document. Finally, step 390 creates a mapit.sfc.weights file 70 for all documents containing the associated subject field codes and weights. Processing will now continue with step 120 of flowchart 100 as depicted on Fig. 2B.

Mapit-all step 120 creates a scores file 74 from a weights file 74. This file has one integer weight from 0 to 99 for every pair of documents in the input document dataset. For example, given documents D1 and D2, corresponding with weight vectors w1 and w2 held in a weights file (such as the word vector weights file 72, or the SFC weights file 70),

corresponding normalization constants n1 and n2 held in a file (created in step 330, and combination function f(w1,w2) defined hereinbelow, mapit-all determines the maximum of a normalized similarity of weight vector W_1 with respect to weight vector W_2 and a normalized similarity of weight vector W_2 with respect to weight vector W_1.

In a related embodiment, a cross-comparison algorithm takes the average of single versus merged claims and merged versus merged claims. For example, to implement the legal concept of patent infringement as applied to sets of patents or patent applications, in a particular embodiment, a similarity matching algorithm treats the exemplar part of a patent claim differently from the dependent parts of the claim. Thus, a kind of "cross-comparison" matching is used, wherein the combined scores for (1) patent A, claim X dependent and independent part(s) vs. patent B, claim Y, independent part and (2) patent A, claim X dependent and independent part(s) vs. patent B, claim Y, dependent and independent part(s), generate an aggregate matching (or similarity) score for patent A, claim X vs. patent B, claim Y.

In cross comparison processing, weights, from either word vector analysis or SFC analysis, are compared from the single file, block 54 of Fig. 2A, and the merged file, block 56 of Fig. 2A. For example, document 1 with weight vectors w1s in the single file and w1m in the merged file is cross compared with document 2, having weight vectors w2s in the single file and w2m in the merged file. The cross comparison score is basically an average of two combination functions of single and merged weights, computed according to formula (1):

$$f'(w1,w2)=(f(w1s,w2m)+f(w1m,w2m))/2.0. \qquad (1)$$

Following step 120 of Fig. 2B, mapit-all-by-patent step 122 aggregates claim level scores to the patent level producing a file containing these patent scores 76. In a preferable embodiment the score for patent p1 versus patent p2 is the top scoring pair of any claim from p1 against any claim from p2. Mapit-all-by-patent implements a "search patents by best claim" function in the preferable embodiment of the invention. The other patent level search, "search patents by all claims" is achieved by performing a regular query against the justclaims data set (i.e., all justclaims files 50 of patents in the associated search set) instead of the top scoring claim in the justclaims data set.

Referring again to Fig. 2B, mapit-top-scores step 126 writes the top N scores to an ASCII format file 82. The rationale underlying this step is that large data file search time is expensive in terms of computing resources. Therefore, in a preferable embodiment, the system precomputes a manageable size score which is the system's "best guess" at what will be of interest to the user. In a preferable embodiment this is implemented by performing a mapit-extract step (i.e., step 300 of Fig. 4A), sorting the resulting file by score, determining the value of

the Nth (i.e., lowest) score, doing a restricted mapit-extract step only down to that Nth score level, and sorting again.

Mapit-score-range step 128 takes as its input the file 82 created in step 126, and calculates the minimum and maximum scores for both word vector analysis and SFC type scores. It then writes this information to a standard output (STDOUT) which has typically been redirected to a convenient file 84.

Following step 128, viz2d step 130 produces a two dimensional plot of top scoring claims, where a score indicates the relative similarity between two claims. Scores are based on word vector analysis. Simultaneously, claim information is aggregated to the patent level in order to depict relationships between patents based upon the similarity of their claims. Claim matches are aggregated together to provide a ranking method (based on a voting-type technique, a technique well known to those having ordinary skill in the art). For patents, this is useful in producing "company A vs. company B" type displays.

In a preferable embodiment, after the top matching pair of claims (i.e., the two claims having the most similarity) in the data set is found, the system rounds the score down to the nearest multiple of 5. Call this score X. Next, three regions are defined. The top region is defined as the rounded score to the rounded score +5 (x to x+5). The middle region is defined as the rounded score -5 to the rounded score -1 (x-5 to x-1). The bottom region is defined as the rounded score -15 to the rounded score -6 (x-15 to x-6).

For each pair of patents p1 and p2, a comparison is drawn for each claim from p1 against each claim from p2 and the following number of points are added to p1 versus p2. Ten is added if the two claims score in the top range. Five is added if the two claims score in the middle range. One is added if the two claims score in the bottom range. Zero is added if the score falls below the bottom range and it is not plotted. Claims falling into each range may be distinguished on the two-dimensional plot through any appropriate identifier such as color coding or symbols. For example, the top, middle and bottom ranges may be plotted with points having colors red, blue and gray, respectively.

All claims at or above the bottom range are plotted and the top ten patent pairs, as scored by the method described hereinabove, are labeled on the graph. The graph is written to a graphs/viz2d.* file 86 and the top ten patent pairs are also written to a separate graphs/viz2d.*.top10 file 88. In a preferable embodiment, step 130 employs the UNIX utility gnuplot to generate a postscript plot and then uses the gs UNIX utility to convert the output of the prior step to a ppm file, which is then converted to a gif file using ppm as is well known by those having ordinary skill in the art. An example of such a plot is provided in Fig. 8B.

Returning again to Fig. 2B, viz3d step 132 produces a three dimensional plot of top scoring claims while simultaneously aggregating

claim information to the patent level. Its functioning is much the same as that of step viz2d 130. However, it gives a 3-D projection of the results and does not label the top ten matches on the graph. An example of such a plot is provided in Fig. 8C.

Finally, viz-compare step 134 produces a cluster plot (also referred to as a "scatter plot" of all the claim pairs from a data set. In contrast to viz2d step 130 and viz3d step 132, wherein the x-axis is one claim number, the y-axis is another claim number, and a dot is plotted if that pair of documents scores above the bottom threshold, the method of viz-compare is that the x-axis represents a wordvec score, the y-axis represents an SFC score, and a dot is plotted if there exists a pair of claims having the corresponding wordvec and SFC scores. An example of such a plot is provided in Fig. 8A.

The scores plotted in Figs. 8A-8C are used to identify documents most closely or proximally related; i.e., "proximity scores". However, such scores may also be plotted to identify those documents that are most different or distally related; i.e., "distal scores". An example of the latter may be seen in Fig. 8D (discussed below). Such distal scores may also be plotted in the charts of Figs. 8A-8C. As such, scores plotted to show relationships among documents are more generally referred to herein as utility measures.

In an alternative embodiment, a user of the system may select which plot type(s) desired by selectively engaging steps 130, 132 and/or 134 of Fig. 2B.

Having detailed the off-line processing component, we now turn to the on-line concept query processing aspect of the invention.

On Line Concept Query Processing

In a concept query, as contrasted to a document query, the user has entered an arbitrary text string (which may be user-originated or copied from a portion or all of a document) which the system must match against the body of known documents to be analyzed (e.g., the dataset). Thus, many of the off-line processing steps described above must be performed against the on-line entered string to get the text into a usable format. Flowchart 401 of Fig. 5 depicts the online query processing. Initially, a user's query input to the system is written to an ASCII formatted file 82, pursuant to step 400.

Actual query processing is handled through a shell script, pursuant to mapit-query step 410. Mapit-query step 410 performs the following processing steps: 1) Build a claimlist, 2) preprocess, 3) tf, 4) tfidf0, and 5) tfidf-all. These are identical in function to the following steps in the off-line claims processing section described hereinabove: 1) "build claimlist" function of make-claims step 106 in Fig. 2A. The system builds a claimlist data structure 84 from the user's query stored as an ASCII format file 82, pursuant to step 400. The

resulting structure is the same in format as claimlist data structure 12 of Fig. 2. 2) preprocess step 108 in Fig. 2A, 3) tf step 300 in Fig. 4A, 4) tfidf0 step 310 in Fig. 4A, and tfidf-all step 320 in Fig. 4A. The output of mapit-query is a set of scores from analysis of the user's query, which are written to a query weight file 86.

Following step 410, mapit-process-query step 420 builds a full score file 90 from input query weight file 86, containing the weights of word stems in the user's query, and a document weights file 88, produced during the off-line processing of the document database as described hereinabove, containing the weights of word stems in the document database. The full score file possesses one integer weight 0-99 for every document in a body or set of documents being processed.

Mapit-all step 120 creates a scores file 74 from a weights file 74. This file has one integer weight from 0 to 99 for every pair of documents in the input document dataset. For example, given documents D1 and D2, corresponding with weight vectors w1 and w2 held in a weights file (such as the word vector weights file 72, or the SFC weights file 70), corresponding normalization constants n1 and n2 held in a file (created in step 330, and combination function f(w1,w2) defined hereinbelow, mapit-all determines the maximum of a normalized similarity of weight vector W_1 with respect to weight vector W_2 and a normalized similarity of weight vector W_2 with respect to weight vector W_1.

Finally, in step 430 the results are converted into a "stars" representation. One star is given for any document with a score greater than zero. An additional star is given for every twenty points in a documents score. The stars are displayed to the user as a representation of the score.

In applications where a response time is critical and/or a large set of documents requires searching, (e.g., based on weights and scores), well-known enhancements may be added to the system to increase processing speed such as use of index access method or other techniques to optimize fast storage and retrieval of data as are well known to persons of ordinary skill in the art.

In a further embodiment, documents are processed according to the off-line processing method described hereinabove to the point where plots are generated in accordance with steps 130-134 of Fig. 2B.

Off-line Processing of Non-structured (Generic) Documents

Flow chart 501 of Figs. 6A and 6B describe off-line processing of non-structured or generic documents (e.g., technical publications, non-structured portions of structured documents (e.g., abstract and detailed description of patent), etc.). For the purposes of this discussion, Figs. 6A and 6B are compared to Figs. 2A and 2B to highlight the differences between off-line processing of structured documents, and off-line processing of generic documents. Off-line generic document processing

begins with creating a file containing the text of the entire document, pursuant to step 502. Input to this file is a text formatted file 11 containing documents in the subject search set. Output is a text file 51. Following step 502, a file 53 is created pursuant to step 504, which contains the full directory path name for each document in file 51. Comparing off-line generic document processing with off-line structured document processing indicates that there is no analog to the make-claims step 106 in generic document processing. Furthermore, single file 54 and merged file 56 outputs of the structured document processing make-claims step do not exist in the generic document processing.

Processing continues with preprocess step 508. Preprocess step 508 is virtually identical to preprocess step 108 (Fig. 2A) of off-line structured claim processing. Preprocessing is described in detail in Fig. 3 as well as hereinabove. Processing continues with step build-hash 512 (build-claimlist step 110 is omitted from flow chart 501), which creates hashed files 59. These files are a subset of the files 60 created in structured document processing and include: 1) A mapping from a claim number to a unique document index, representative of each document being analyzed; 2) A mapping from a claim number to the full directory path to the text of that claim; 3) A mapping from a claim number to the first 150 characters of the claimThe fix-patentlist step 114 of structured document processing (Fig. 2A) is omitted in the generic document processing of Fig. 6A. The generic processing continues with mapit-process-generic step 516. The mapit-process-generic step is virtually identical to the mapit-process step 116 of structured claim processing. Mapit-process is described in detail in Figs. a and 4B and herein above. The output of mapit-process-generic step 516 includes an SFC input file 61 and a mapit.sfc.weights file 63. These files are identical to files 60 and 62, respectively, of Fig. 2A. Off-line generic processing continues on Fig. 6B with mapit-all step 520 which builds a scores file 75 from a weights file 63. Since there are no structured elements such as claims in generic documents, there is no equivalent to the mapit-all by patent step 122. So generic document processing continues with retrofit-sfc step 520, which functions as its counterpart retrofit-SFC step 124 in Fig. 2B. Retrofit-SFC step 520 applies word vector analysis information to the SFC weighted scores, producing a new SFC score file 81 and saving the original information in an original file 79. The processing continues with mapit-top-scores step 526 which creates a file 83 of top scores. Finally, mapit-score-range step 528 computes the minimum and maximum scores and writes them into file 85. This information may be output as an individual data file using conventional means.

"Generic" documents in this context may include claims treated as a generic document (i.e., without parsing) compared with other portions of a patent (e.g., summary, abstract, detailed description, etc.).

In an alternative embodiment, it is contemplated that plot generation including two dimensional, three dimensional and cluster will

be available with generic document processing. This feature will be enabled in accordance with the methodolgy discussed above for structured document processing.

Claim Parsing According to a Specific Embodiment

Fig. 7A shows a Flow chart 650 with a method of parsing claims according to a specific embodiment of the invention. In a preferred embodiment, the method of flow chart 650 is employed by step 106 of Fig. 2A to create files 54 and 56. The input to the claims parsing process is a single file containing a set of all claims from a patent (e.g., justclaims 50). The output is a single file and a merged file for each claim. The single file will contain only the body of a single claim. The merged file will contain the body of the single claim in addition to the transitive closure of all claims referenced therein. These files are identical to files 54 and 56, respectively, of Fig. 2A.

The process reads claims from the justclaims input file 50, one line at a time in the "get the next line" step 600. The system then determines if the line read in step 600 is the start of a new claim in step 602. New claims are indicated to the system by a fresh line starting with a claim number followed by a period, a space and the claim text. Claim numbers must be sequential and begin with the number 1. If the system detects the beginning of a new claim then the system will add the current claim that it had been processing to the claim list file 12 (list of document names) in step 604. Otherwise, or in any event, in step 606 the system appends the current line read in from the file to the current or new claim body. Next, the system will determine whether another claim is referenced in the current line read in from the file, pursuant to step 608. If a reference is indicated, the system will read in the next line from the input file in step 610. This is done in case the reference crosses a line boundary. The system will also try to identify claim references in step 610. Note that there are two simplifying assumptions. Number one, claim references never run more than two lines. Number two, a new claim reference is never detected on the second line which continues to the third line.

In the alternative, or in any event, the system tokenizes the line saving the tokens into an array pursuant to step 612. All matter in the line up to the word claim is discarded. For example, in the line, "5. The method of claim 1", this step would eliminate all text prior to the word "claim", i.e., "5. The method of". Tokens are not split based upon punctuation because it creates extra tokens. Ending, or trailing punctuation is removed from the end of words in step 614. The last word in the line is saved in a variable "last_word" in step 616 to facilitate the check for the words "preceeding" or "previous" in step 622 of Fig. 7B.

Having tokenized the line into words, the system will now invoke process words in step 618, as described below, to look for

references to other claims within the line. Upon completing step 618, a determination is made as to whether there are any more lines in the input file (i.e., just claims 50) in step 619. If yes, control returns to step 600 to process the next line in the current or a new claim. If not, parsing is complete for the set of claims of the subject patent and parsing processing stops (unless another patent is to be processed).

Referring to flow chart 652 in Fig. 7B, words in the array are processed serially beginning with the "get next word" step 620, which fetches a current word. The system checks for the existence of the word "previous" or "proceeding" in step 622. If the "last_word" was previous or proceeding, then the system understands this to indicate that it should add all claims including this one to claim list file 12 in step 624. In the alternative, processing proceeds with the system checking for a plain (i.e., arabic) number in step 626. If the system detects a plain number then the system understands this to indicate that a new claim has been found, and that the current claim should be added to claim list file 12, pursuant to step 628. In the alternative, the system next checks the current word for an "or" an "and" or an "inclusive" in step 630. If the system detects the presence of either of these three words this word is skipped and no processing is done in step 632. In the alternative, processing proceeds with examining the current word for a hyphenated range pursuant to step 634 (for example, claims 4-19). If the system detects the presence of a hyphenated range, the system adds the claims in the range to claim list file 12, in accordance with step 636. In the alternative, processing proceeds to check for the existence of a range delimited by the words "to" or "through" in step 638. If the system detects a "to" or a "through" delimited range, the system adds the claims in the range to claim list file 12, pursuant to step 640. In the alternative, the system detects the condition that there is nothing more to reference. At this point, the system has detected that this is the end of the claim reference. Processing continues with the system searching for another claim reference within the subject line, pursuant to step 642. Next, in step 644, the current word is saved in the "last_word" variable. The system next determines whether there are any more words in the subject line being processed, pursuant to step 645. If not, control returns to step 619 of flow chart 651. Otherwise, in preparation for another iteration through the loop, control flows back to the beginning of the process-words step, where the "get next word" step 620 is executed to process the next word in the set of words.

Ultimately, when all of the words in a line are reached, control flows to step 619 in Fig. 7A, which detects if the last line of the claim has been processed. If so, processing halts for this claim. Otherwise, control returns to the get-next-line step 600.

Graphical Display and Visualization of Analysis Results

Figs 8A-8D illustrate examples of formats in which to display

and analyze document data as provided by a particular embodiment of the invention.

Typical clustering techniques, known in the art, represent documents as points in an n-dimensional display, wherein each point corresponds to a single document and each dimension corresponds to a document attribute. These clusters are then typically displayed as graphical images where related documents are indicated by spatial proximity (sometimes further distinguished by color or shape). Examples of this sort of clustering include the "Themescape" type displays from Battelle, a corporation with headquarters in Columbus, Ohio.

Contrastingly, according to the invention, clustering is performed using a single point in n-dimensional space to represent a pair of documents, rather than a single document. Each dimension represents a separate metric measuring the similarity of the two documents. By using different sets of orthogonal metrics, clustering of underlying documents can be performed in different ways to highlight different features of the overall collection.

A set of metrics can be selected for display. For example, Fig. 8A depicts two orthogonal similarity metrics which scores: thematic similarity 702 (in the form of semantic thread score or SFC-type score) identifying documents about the same topic even if they use different terminology, and syntactic similarity 704 (in the form of word vector score) which identifies documents that use the same terms and phrases. These metrics may employ differing matching techniques. For example, a subject field code (SFC) vector technique may be combined with a space metric based on TF.IDF weighted term occurrences.

Preferably, thematic similarity is determined employing SFC techniques described by Dr. Elizabeth Liddy in the above-referenced articles. Further, syntactic similarity is determined through word-vector analysis using TF.IDF techniques, which are well known to those having ordinary skill in the art and more further described in the Salton reference. In a preferable embodiment this set of metrics is displayed visually as an x-y scatter plot, as in Fig. 8A, although clusters can be displayed within larger dimension sets by using additional graphical attributes such as 3D position, size, shape, and color.

Many systems use a combination algorithm to collapse multiple similarity measures into a single value. According to the invention, the individual similarity components in the visual display are retained, allowing the user to interpret the multiple dimensions directly. For example, for certain patent applications, it may be useful to identify document pairs that are similar across both dimensions, while for other applications it may be more important to identify cases where the two similarity scores differ. The user can interactively explore the visualization by using a mouse or other input device to indicate either a single point (a single pair of documents) or regions of points (a cluster of document pairs). The documents represented by these points can then be

displayed, either by presenting full text or by presenting identifying attributes such as title and author. The ability to cluster and display documents using multiple similarity measures simultaneously would be lost if everything were collapsed to a single score.

Scatter Diagram:

Fig. 8A illustrates a scatter plot for drawing inferences from A vs. B types of analyses according to the method described above in the viz-compare step 134 of Fig. 2B. A collection of documents may be split into two sets, for example, patents from Company A and patents from Company B. Paired Proximity scores are developed, using the method described hereinabove, one score for every document in set A against every document in set B, and the other score for every document in set B against every document in set A.

In the scatter plot, the x-axis represents relative similarity according to a syntactic or word vector based score. The y-axis depicts relative similarity based on a conceptual or semantic thread based score. In a split dataset, each document from the first dataset is compared against the documents of the other dataset, resulting in a score represented by a point in the space defined by the syntactic and semantic axes. Documents which are highly similar according to word vector based analysis will appear farthest to the right on the plot. Documents having the highest similarity according to a semantic based analysis will appear at the top of the plot. Documents having the greatest similarity to one another based upon both word vector and semantic thread score will appear in the upper right hand corner of the plot. Documents having the least amount of similarity according to both word vector and semantic scores will appear in the lower left hand corner of the plot.

In a related embodiment, the highest proximity scores for each document in set A against entire set B, and highest proximity scores for each document in set B against entire set A are determined.

In a related embodiment, zooming-in or zooming-out in a scatter plot increases or decreases the resolution and range/domain of the plot.

2-D Diagram:

Fig. 8B illustrates a 2D visualization of an analysis conducted on two sets of patents according to the method described in the viz2d step 130 of Fig. 2B. In the 2-D plot, the x-axis exhibits the patents in the dataset as monotonically increasing sequence of patent numbers. The y-axis is identical to the x-axis. Clusters of the most similar patents within the dataset are plotted on the graph. Clusters with scores falling within the 95 to 100 range are plotted with a square. Clusters with a score falling within the 90 to 94 range are plotted with a cross. Clusters with a score falling within the 80 to 89 range are plotted with a circle.

In a related embodiment, color is added to the 2D, orthogonal similarity plot according to various criteria. For example, if the user types in a search concept "digital image segmentation and edge detection," patent components shown in the plot will change color (or some other display appearance attribute) according to the strength of presence of this concept in the data. This may be carried out with an overlay plot applied to the 2-D diagram.

3-D Diagram:

Fig. 8C illustrates a 3D visualization of an analysis conducted on two sets of patents according to the method described in the viz3d step 132 of Fig. 2B. The 3-D diagram depicts the same information as the 2-D diagram only in a three dimensional format. The x-axis and y-axis both are delienated by monotonically increasing numbers of the patents in the dataset. The z-axis represents a ranged degree of similarity of the patents. Scores based on the similarity of clusters of patents are plotted in the 3-D framework with the same graphical representations as in the 2-D plot described hereinabove. (i.e., scores within the 95 to 100 range are depicted as a square; scores within the 90 to 94 range are depicted with a cross; scores falling within the 80 to 89 range are depicted by a circle).

S-Curve Diagram:

Fig. 8D illustrates an S-curve plot for drawing inferences from A vs. B types of analyses. In this method of displaying data analysis results, documents from dataset A are plotted on the left hand side with low proximity scores having negative values with large absolute values, and where documents from dataset B are plotted on the right hand side with low proximity scores having positive values with large absolute values. In other words, plot (score - 1.0) for set A documents and (1.0 - score) for set B documents, then sort and plot to yield an S-shaped curve).

Fig. 8E illustrates the steps to produce the S-curve. The process depicted in the flow chart 801 begins with the generation of all scores either term or concept from a claim level data set A versus data set B analysis 850. For example, the patents from Company A compared with the patents from Company B on a claim by claim basis. These scores are in the range of 0.0 to 1.0. Next, in step 852, all claims are sequentially numbered such that the first claim from Company A is 1 and the last claim from Company B is n and all claims from A precede all claims from B. In step 854, for each claim index I from Company A find the closest claim from Company B and record the pair (I, S-1.0), for S is the similarity score of A compared with B. Next, in step 856, for each claim index I from Company B find the closest claim from company A and record the pair (I, 1.0-S) where S is the similarity score of A compared to B. Finally, in step 858, sort all pairs in increasing order of second coordinate and

display on a plot where the x-axis represents the claim index and the y-axis represents the claim score.

The result is a plot in the form of an S-curve where the bottom part of the S represents claims unique to company A; the middle part represents claims with possible overlaps between the two companies, and the top part represents claims unique to Company B.

In a related embodiment, the S-curve method of displaying data is extended to analyses wherein additional documents are added to sets A and/or B and reanalyzed. The resulting graph is overlaid on top of the original graph. This permits the user to track changes over time, for example where changes in the shape of an S-curve of patent portfolios represent changes in the technology holdings of one company relative to the other.

Techniques for Analysis of Documents

Screens (also referred to as "pages" herein) and automated tools incorporated in a specific embodiment of the invention enables a user to perform detailed study on analysis results of the system. Fig. 9A, for example, depicts a representative sign-on screen for a user according to the invention. Screens are produced using the NetScape NetBrowser interface to the worldwide web. The reader of ordinary skill in the art will appreciate that other web browsers and other programs may be used as a user interface to the patent analysis aspect of this invention. The user enters a user I.D. and password in the screen depicted by Fig. 9A to sign-on to the system described herein. After the password and I.D. have been authenticated, in one embodiment of the present invention, a dataset representing a portion of the U.S. patent database (e.g., over 2 million patents) is automatically selected. In another embodiment, it is necessary to choose an initial dataset to analyze. Exemplary dataset types include: Portfolio Analytics, Custom Canvas, Products, World Patents and Industry Verticals.

Portfolio Analytics contains patent datasets (i.e., sets of patents). There are two types of patent sets: single and split. Single patent sets contain all patents together in one group. All search and analysis functions are applied to all of the patents and claims in the patent set. In contrast, split sets contain two groups of patents. These two independent patent groups are measured against each other during comparative analysis. For example, if a split set contains information about company A in one patent group and company B in another group, then a claim query or patent query with a patent from the company A group will display the company A item versus a company B item. An exemplary screen shot of dataset selection is provided in Fig. 9B.

The remaining exemplary datasets include Custom Canvas (which will contain user-defined sets), Products (which will contain product datasets for patent versus product analysis), World Patents (which will contain patent sets grouped by geographical region) and Industry Verticals

(which will contain industry-specific patent sets).

Figs. 9C - 9K depict representative screens in accordance with the performing of a concept query as described herein above. A concept query entry screen 900 depicted in Fig. 9C enables the user to enter in English text, a description of a concept which the system will search for in the database of patents. The concept entry screen has fields which enable the user to specify a job I.D. for billing purposes and to search sections by abstracts or claims and also to control the order of sorting. Further, screen 900 provides a NASA Thesaurus link 902 which, when clicked upon, launches a Netscape window with the index of the NASA Thesaurus. A term found in the thesaurus may be included in the query by copying and pasting or simply typing the word into query box 904.

Screen 900 also includes a search selection box 906 which is used to define the scope of a query and results. The options for box 906 include "claims," "patents (best claim)" and "patents (all claims)." In the "claims" option, the system searches each individual claim in the selected dataset and returns a results list ranked by claim score. The results list, as shown in the screen of Fig. 9E, displays patent information 916, 923 as well as claim information 918, including a preview of the claim text 920.

In the "patents (best claim)" option, the system searches each individual claim in the selected dataset and returns a results list ranked by patent, where the patent score is based on the score of the highest ranked claim in the patent. The results list displays patent information.

In the "patents (all claims)" option, the system searches the combined (i.e., all) claims for each patent and returns a results list ranked by patent, where the patent score is based on a score for all the claims in the patent. The results list, as shown in Fig. 9F, displays patent information 926, 928.

Referring again to Fig. 9C, clicking on Analyze Query button 903 produces a concept query review screen 909 of Fig. 9D, which depicts the results of the stemming operations described hereinabove as applied to the user's concept query which has been entered in screen 900 of Fig. 9C. For each stemmed word and phrase entered in the concept query, the concept query review screen indicates the number of claims 912 and patents 914 containing each word or phrase. By clicking a "show results" button 915 on screen 909, the user may go to a "concept query results" screen depicted in Fig. 9E (for a "claims" search) or 9F (for a "patents (all claims)" search).

Referring to Fig. 9E, a concept query results screen 917 provides the results of a user's "claims" search as applied to the database of patents. For the representative query depicted in box 919 of Fig. 9E, the results are provided in a list ranked by claim score. The Relevance level 921 of any given claim is indicated by the number of stars from one (worst) to five (best). A user may click on a rank number 922 to move to a screen showing a side-by-side comparison of the associated claim

and the original query (Fig. 9I). Additionally, a user may click on a patent number 916 to move to a screen showing the full text of the patent (Fig. 9K) and on a claim number 918 to move to a screen showing the full text of the claim (Fig. 9J). These linked screens are described in more detail below.

Screen 917 (like many screens described below) contains a number of "links" to other screens in forms which include rank numbers, patent numbers and claim numbers. These inter-screen links may be achieved using HTML (e.g., via hyperlinks) or any other conventional method known to those having ordinary skill in the art. Such links provide a convenient and well-known mechanism to "navigate" between screens containing information desired by a user. In a preferred embodiment of the invention, clicking on a claim number, patent number or rank number in any screen in which such numbers represent links will call a "viewer" function, which loads the relevant text described above into a separate window.

More specifically, clicking on a rank number 922 results in a link to a viewer side-by-side comparison screen in the form of screen 970 of Fig. 9I. As shown therein, the left half of the screen 972 contains the full text of a concept query while the right side of the screen 974 includes the title, assignee, patent number, and full text of the subject claim. According to one embodiment, if a subject claim refers to a previous claim (i.e., it is a dependent claim), all the claims referenced, either directly or indirectly (i.e., the transitive closure of the subject claim) will be shown in the order referenced. According to another embodiment, if a subject claim refers to a first previous claim, the first previous claim number will be in the form of a link embedded in the text of the subject claim. This link will be to a screen containing the text of the first previous claim. In like fashion, if the first previous claim refers to a second previous claim, a second link (in the form of the second previous claim number) will be embedded in the text of the first previous claim to a screen containing the text of the second previous claim. This daisy chain of links continues until the family of claims is traced back to the associated independent claim(s).

Referring again to Fig. 9I, patent number 986 in Fig. 9I functions as a link to a screen containing the full text of the subject patent. In addition, highlighting controls 976-982 are provided in this screen. Such controls allow a user to highlight text in any of the text areas displayed using two colors. Words or phrases are inserted into boxes 976 and 980, and desired colors are chosen in boxes 978 and 982, respectively. Upon clicking update button 984, the desired words and phrases in all of the text windows will be highlighted using the colors indicated 978, 982 for each text box 976, 980.

Referring back to Fig. 9E, clicking on a claim number 918 links to a claim viewer screen in the form of screen 980 of Fig. 9J. As shown therein, this screen is essentially the same as screen 970 (like

reference numbers refer to like features) without left-half portion 972. Again, if a subject claim refers to a previous claim, in one embodiment the transitive closure of the claim (i.e., all claims referenced either directly or indirectly) shall be shown in the order referenced. Alternatively, in another embodiment the subject claim shall include in its text the claim number of the directly referenced claim(s) in the form of a link to another screen or screen(s) containing the text of the referenced claim(s).

Referring back to Fig. 9E, clicking on a patent number 916 links to a patent viewer screen in the form of screen 990 of Fig. 9K. As shown therein, this screen has many of the same features as screens 970 and 980 (like reference numbers refer to like features). In addition, screen 990 includes window 992 which may contain the full text of the subject patent. Alternatively, in another embodiment, window 992 may display an abbreviated disclosure including the patent title, assignee, bibliographic information, abstract and full text of claims. Whether a full text or abbreviated disclosure of the subject patent is provided when clicking on a patent number may be determined by the type of dataset being searched; not all datasets will necessarily contain full text documents.

In addition to the "claims" based results shown in Fig. 9E, a concept query results screen 925 in Fig. 9F gives the results of a user's "patents (all claims)" search as applied to the database of patents. In the representative query depicted in Fig. 9F, the patents are listed in order of decreasing relevance to the user's concept query (shown in block 930). Patents are ranked in numerical order and a patent number 926 is given along with a title and an assignee 928. Next, the user may by clicking on a patent number 926, move to a screen showing the full text of the patent (in the same form of Fig. 9K). In an alternative embodiment, the user may by clicking on patent number 926, move to a screen which provides an "abbreviated" disclosure of the subject patent. This abbreviated version may be in the form described above in connection with Fig. 9K or, alternatively, in the form of screen 950 of Fig. 9G. Specifically, window 951 of screen 950 provides an abbreviated section describing the inventors, assignees, filing dates, categories and classes of the subject patent. Also included is a table of U.S. references, abstract, and the claims of the patent (not shown). As noted above, whether a full text or abbreviated disclosure of the subject patent is provided when clicking on a patent number may be determined by the type of dataset being searched; not all datasets will necessarily contain full text documents. (This is also true for patent and claim queries, described below.) Window 951 also includes a View Image link 952 which, when clicked, will launch a new Netscape window from a particular server site (e.g., http://my_patent_site.com) containing images and will load a scanned image of the subject patent into the window.

Screen 925 of Fig. 9F also includes a Modify Query link 927 which may be clicked on to return to the original query.

Referring again to Fig. 9F, the user may also click on a rank number 932 and move to a patent viewer side-by-side screen 960 as depicted in Fig. 9H. The patent viewer screen 960 of Fig. 9H enables the user to have a side-by-side comparison of the concept query entered and the text of various patents which match the concept query according to the system. The full text of these documents is presented simultaneously on a computer display, enabling the user to interactively explore a comparison of the two documents. Alternatively, a subset of the text may be provided that includes the abstract, claims and/or bibliographic information. The format, as noted above, may be determined by the type of dataset being searched. In addition, screen 960 includes highlighting controls 976-982 like those of Fig. 9I.

Figs. 10A, 10B, 10C and 10D depict representative screens in accordance with the performing of a patent query as described hereinabove. The patent query allows the user to draw comparisons between a single patent and all other patents in the dataset. If the dataset is a single dataset (i.e., not a split dataset) the patent query will compare the selected patent to all of the patents in the selected dataset. If the selected dataset is a Split dataset (i.e., having at least two data groups), the selected patent is compared just to the group of patents that it is not in.

A patent query entry screen 1000 depicted in Fig. 10A enables the user to enter the number of a patent contained in the database of patents in block 1002. The system will analyze all members of the database of patents against the patent entered. (However, when "Filter Out Claims" selector 1005 is checked, the system will not compare claims from the same patent.) Like concept query, the patent query screen has a search field 1004 which enables the user to select search processing for "patents (all claims)" or "patents (best claim)." In "patents (best claim)" processing, the patent is compared to each individual claim in the selected dataset (for a single dataset), or to each individual claim in the data group not containing the selected patent (for a split dataset), and returns a results list ranked by patent. The patent score is the score of the highest ranked claim in the patent. The results list displays patent information and has an option to view a listing of all the ranked claim pairs for any patent in the results list.

In "patents (all claims)" processing, the patent is compared to all of the combined claims for each patent in the selected dataset (for a single dataset), or to an amalgamation of claims for each patent in the data group which the selected patent does not belong to (for a split dataset), and returns a results list that ranks each matching patent based on a score for all the claims in the patent. The results list displays patent information and has an option to view a listing of all the ranked claim pairs for any patent in the results list.

Returning to Fig. 10A, clicking on Show Results icon 1003 displays the query patent number, title and assignee information at the

top of a results screen 1010, as shown in Fig. 10B.

The patent query results screen 1010 depicted in Fig. 10B gives the results of the user's search as applied to the database of patents. In the representative query depicted in Fig. 10B, the patents are listed in order of decreasing relevance to the user's query. Patents are ranked in numerical order and the patent number 1012 is given along with the title and an assignee 1014. As shown in Fig. 10B, a patent query generates two scores for each result; a Phrase Score 1018 and a Theme Score 1020. Phrase Score 1018, generated from word-vector analysis, measures similarities based upon words and phrases in claims. Theme Score 1020, generated from semantic thread analysis, measures similarities based upon topical themes and concepts. The score used to sort is displayed in bold.

Screen 1010 provides several navigational links between screens. For example, by clicking on a patent number 1012 the user may move to a screen which displays the entire text of the patent (in the same form as shown in Fig. 9K). Alternatively, the user may click on a "view claims" link 1016 to arrive at a claims comparison screen 1030 depicted in Fig. 10C. Claims comparison screen 1030 permits the user to identify matching claim pairs between the two patents at issue.

Referring to Fig. 10C, screen 1030 includes query patent number 1032 and results patent number 1034. These patent numbers form links to screens displaying the entire text of the patents (in the same form as shown in Fig. 9K). The matching claim pairs for the two patents are listed in rank order; e.g., rank 1 (claims 20, 2) and rank 2 (claims 21, 2). Corresponding claim numbers 1036 form links to screens displaying the entire text of the claims (in the same form as shown in Fig. 9J). Further, rank numbers 1038 form links to a side-by-side viewer screen in the form of screen 1040 of Fig. 10D.

As shown in Fig. 10D, screen 1040 has many of the same features as screen 990 of Fig. 9K (like reference numbers refer to like features). Notably, screen 1040 includes windows 992 which may contain the full text of the subject patents. As shown therein, the patent viewer screen enables the user to have a side-by-side comparison of the two patents. The full text of these documents presented simultaneously on a computer display enables a user to interactively explore a comparison of the two documents. Alternatively, in another embodiment, windows 992 may display an abbreviated disclosure including the patent title, assignee, bibliographic information, abstract and full text of claims. As noted above, the type of patent information provided may be determined by the type of dataset being searched.

Additionally, the user may click on a rank number 1013 of Fig.10B, which also links to the side-by-side viewer screen 1040 depicted in Fig. 10D.

Figs. 11A, 11B and 11C depict representative screens in accordance with the performing of a claim query as described hereinabove.

The claim query allows the user to draw comparisons between a single claim and all other claims in the dataset. If the dataset is a single dataset, the claim query will compare the selected claim to all of the claims in the selected dataset. If the selected dataset is a Split set (having two data groups), the selected claim is compared just to the group of claims that it is not in.

Claim query entry screen 1102 depicted in Fig. 11A enables the user to enter the number of a patent and a claim contained in the database via data entry blocks 1104 and 1106, respectively. The system will analyze all members of the database against the claim entered. A user who is unsure of the correct claim number to enter, may, after entering the patent number, click a "view claims" icon 1108, which will display the full text of the claims as shown in Fig. 11B. Screen 1120 of Fig. 11B displays the entire text of the claims for the patent corresponding to the patent number entered. The user can scroll through the claims until the desired claim is found.

Referring again to Fig. 11A, the claims query function, as specified in block 1110, will compare a selected claim to each individual claim in the selected dataset (for a single set) or to each individual claim in the data group to which the selected claim does not belong (for a split set). It returns a results list ranked by claim.

Once the user has entered the desired claim and selected a "show results" icon 1112, the system responds with matching claims in ranked order in screen 1130 of Fig. 11C. As shown in Fig. 10C, a claim query generates two scores for each result; a Phrase Score 1132 and a Theme Score 1134. These scores have the same meaning as Phrase score 1018 and Theme score 1020, respectively; which are described in connection with Fig. 10B. Screen 1130 also provides query patent number 1136 and resulting patent number 1137, along with corresponding claim numbers 1138 and 1139, respectively. These patent numbers form links to screens displaying the entire text of the associated patents (in the same form as shown in Fig. 9K). The corresponding claim numbers 1138, 1139 form links to screens displaying the entire text of the claims (in the same form as shown in Fig. 9J).

In addition, a user may click on a hyperlink rank indicator 1140 to perform a side-by-side comparison of the claims, resulting in a side-by-side viewer screen in the form of screen 1150 of Fig. 11D. As shown in Fig. 11D, screen 1150 has many of the same features as screen 980 of Fig. 9J (like reference numbers refer to like features). Each window 1152 and 1154 displays the title, assignee, patent number and full text of the matching claims. Like Fig, 9J, if a subject claim refers to a previous claim, in one embodiment the transitive closure of the claim (i.e., all claims referenced either directly or indirectly) shall be shown in the order referenced. Alternatively, in another embodiment the subject claim shall embed in its text the claim number of the directly referenced claim(s) in the form of a link to another screen or screen(s) containing

the text of the referenced claim(s).

Alternatively, referring again to Fig. 11C, the user can click on "view overlay plot" icon 1142 to view highlights of all the match points of the results over the top of a cluster plot. Fig. 11F depicts the steps in producing the overlay plot. First, as depicted by step 1102 of flow chart 1101, generate the basic cluster plot for an entire data set by offline processing as described hereinabove. Next, according to step 1104, run a claim query and generate score files, both term and concept scores, for each document in the data set against the claim (online) query. Finally, in step 1106, for each match against the claim i, plot ST(i)+e, SC(i)+e on the cluster plot in a contrasting color to the original cluster plot; where in ST(i) equals the term score for document i, SC(i) equals the concept score for document(i), e equals a random epsilon value for spreading. The result is that the dots on the full cluster plot that correspond to the claim query are highlighted.

Figs. 12A and 12B depict representative screens in accordance with the performing of a range query. The range query allows the user to view claim pair matches in the dataset by specifying a score range. If the selected dataset is a single dataset the range of every claim in the dataset is compared to every other claim. If the set is a split set, every claim from the first data group will be compared to every claim from the second data group.

The range query entry screen depicted in Fig. 12A enables the user to enter a start value and end value for a phrase score and a theme score and then to select which score is to be used by the system in order to rank results.

The system ranks the results in the range query as depicted in Fig. 12B. The results are listed by patent number, title, assignee information and the number of lines of each claim. By clicking on the rank number, the user can view a side-by-side comparison of the two claims in the viewer (in the same form as Fig. 11D). Otherwise, by clicking on the patent number the viewer can view the full text of the patent in the viewer (in the same form as Fig. 9K). Or, by clicking on the claim to view, the user may view the full text of the claim in the viewer (in the same form as Fig. 9J).

Fig. 12C depicts steps in producing a range query. First, as shown in step 1202, the user views the cluster plot and decides on an area of interest determined by a rectangle. Next, in step 1204, the user enters the ranges for term scores and concepts scores ST_min, ST_max, SC_min, SC_max in accordance with the rectangular region of interest determined in prior step 1202. The result is a result page showing only the matches that have scores in the specified range corresponding to the rectangle of the cluster plot.

The automated highlighting in the user query screen enables the highlighting of documents displayed side-by-side on the same display where any occurrence of words or phrases from one or more predefined lists

are highlighted visually. Automated highlighting may also be used where any occurrence of words or phrases specified by the user (or any words or phrases automatically generated by one or more sets of rules specified by the user) are highlighted visually.

In a related embodiment, the type of highlighting can be varied to indicate the list to which the highlighted word or phrase belongs, the words or phrases can be highlighted only when they occur in both documents.

An alternative embodiment of the present invention is provided in Figs. 13A-13G. Included in this embodiment is a claim selection operation as shown in screen 1310 of Fig. 13B. Specifically, as shown in this figure, hyperlink claim numbers 1312 are provided for each claim identified in this screen. A user may click on one of these claim numbers to view the underlying claim (i.e., each claim number provides a link to a screen displaying the text of the identified claim). Further, the transitive closure of select claims is provided in screens 1320, 1330 and 1340 of Figs. 13C, 13F and 13G, respectively.

While the foregoing is a complete description of a specific embodiment of the invention, various modifications, alternative constructions and equivalents will be apparent to one skilled in the art. Although aspects of the invention are described in terms examples of analyzing and visualizing patent texts, aspects of the invention are applicable to other classes of documents. Therefore, it is not intended that the invention be limited in any way except as defined by the claims.

WHAT IS CLAIMED IS:

1. A method of analyzing and displaying information regarding a plurality of documents, the method comprising the steps of:

generating a set of N different representations of each document, a given representation being designated the i^{th} representation where i is an integer in the range of 1 to N inclusive;

for selected pairs of documents, determining N utility measures, a given utility measure being designated the i^{th} utility measure where i is an integer in the range of 1 to N inclusive, the i^{th} utility measure being based on the respective i^{th} representations of the documents in that pair; and

displaying a scatter plot in an area bounded by N non-parallel axes, a given axis being designated the i^{th} axis where i is an integer in the range of 1 to N inclusive, where each selected pair is represented by a point in N-space having a coordinate along the i^{th} axis equal to the i^{th} utility measure.

2. The method of claim 1 wherein the set of N different representations comprises:

a first representation being a conceptual-level representation; and

a second representation being a term-based representation.

3. The method of claim 2 wherein the utility measure is a proximity score.

4. A method of analyzing and displaying information regarding a plurality of documents, the method comprising the steps of:

generating first and second different representations of each document;

for selected pairs of documents, determining (a) a first utility measure based on the respective first representations of the documents in that pair, and (b) a second utility measure based on the respective second representations of the documents in that pair; and

displaying a scatter plot in an area bounded by first and second non-parallel axes where each selected pair is represented by a point having a first coordinate along the first axis equal to the first utility measure and a second coordinate along the second axis equal to the second utility measure.

5. The method of claim 4 wherein:

the first representation is a conceptual-level representation; and

the second representation is a term-based representation.

6. The method of claim 4 wherein:
the first representation is a subject vector; and
the second representation is a word vector.

7. The method of claim 4 wherein each of the selected pairs consists of a particular document in the plurality of documents and a different respective one of the remaining documents in the plurality of documents.

8. The method of claim 5 wherein the utility measure is a proximity score.

9. The method of claim 4 wherein the documents are publications.

10. The method of claim 4 wherein the documents are articles from journals.

11. The method of claim 4 wherein the documents are attributable to a product.

12. The method of claim 4 wherein the documents are contained in a split dataset for making comparisons between collections of documents.

13. The method of claim 4 wherein the documents are different parts of patents.

14. The method of claim 13 wherein the different parts of patents include claims.

15. The method of claim 13 wherein the different parts of patents include a detailed description.

16. The method of claim 13 wherein the different parts of patents include an abstract.

17. The method of claim 13 wherein the different parts of patents include a summary.

18. The method of claim 13 wherein the different parts of patents include a Background of Invention.

19. A method of analyzing information regarding a plurality of documents, each having a unique document index, the method comprising the steps of:

parsing each document into a plurality of elements;

generating a first representation of each of said elements; and

for selected pairs of documents, comprised of a first document and a second document, determining a first utility measure based on the respective first representation of the plurality of elements for the documents in that pair.

20. The method of claim 19, wherein said plurality of elements are in a hierarchical relationship, further comprising the step of:

displaying a representation of each of said plurality of elements reflecting said hierarchical relationship.

21. The method of claim 19 wherein said elements comprise patent claims.

22. The method of claim 20 wherein said representation is a hypertext link.

23. The method of claim 20 wherein said representation is a depiction of a sequence of said plurality of elements organized to reflect said hierarchical relationship.

24. The method of claim 19, wherein said plurality of elements are in a hierarchical relationship, further comprising the step of:

selecting a particular element from said plurality of elements as a basis for further analysis.

25. The method of claim 19 wherein the parsing step produces a transitive closure of said plurality of elements.

26. The method of claim 19 wherein the elements are claims and the parsing step comprises the steps of:

reading in text;

determining whether a new claim has begun;

tokenizing said text to extract a plurality of tokens;

adding said plurality of tokens to a word list for the claim; and

scanning said tokenized text for tokens which indicate a reference to a different claim.

27. The method of claim 19 further comprising the step of displaying a plot in an area bounded by first and second non-parallel axes where each selected pair is represented by a point having a first coordinate along the first axis and a second coordinate along the second

axis.

28. The method of claim 27 further comprising the steps of:
generating a second representation of each of said elements;
for the selected pairs of documents, determining a second utility measure based on the respective second representation of the plurality of elements for the documents in that pair; and
wherein in the displaying step, the plot is a scatter plot, the first coordinate is equal to the first utility measure and the second coordinate is equal to the second utility measure.

29. The method of claim 19 further comprising the steps of:
generating a second representation of each of said elements;
for the selected pairs of documents, determining a second utility measure based on the respective second representation of the plurality of elements for the documents in that pair.

30. The method of claim 27 further comprising the steps of:
wherein in the displaying step, the plot is a 2 dimensional visualization, the first coordinate is equal to the unique document index of the first document of a pair of documents and the second coordinate is equal to the unique document index of the second member of a pair of documents, and an icon representing the first utility measure is plotted for each pair of documents.

31. The method of claim 19 further comprising the step of displaying a plot in an area bounded by first, second and third non-parallel axes where each selected pair is represented by a point having a first coordinate along the first axis, a second coordinate along the second axis and a third coordinate along the third axis.

32. The method of claim 31 further comprising the steps of:
wherein in the displaying step, the plot is a 3 dimensional visualization, the first coordinate is equal to the unique document index of the first document of a pair of documents and the second coordinate is equal to the unique document index of the second member of a pair of documents, and the third coordinate is equal to the first utility measure, and an icon representing the first utility measure is plotted for each pair of documents.

33. The method of claim 30 wherein said first utility measure is a combination of N utility measures.

34. The method of claim 32 wherein said first utility measure is a combination of N utility measures.

35. The method of claim 28, for an additional document further comprising:

parsing said additional document into a plurality of elements;

generating a first representation of each of said elements from the parsing step;

for selected pairs of documents drawn such that a first member of the pair is the additional document and a second member of the pair is from said plurality of documents, determining a first utility measure based on the respective first representation of the plurality of elements for the documents in that pair;

generating a second representation of each of said elements from the parsing step;

for selected pairs of documents drawn such that a first member of the pair is the additional document and a second member of the pair is from the plurality of documents, determining a second utility measure based on the respective second representation of the plurality of elements for the documents in that pair; and

wherein in the displaying step, the plot is a scatter plot, generating an overlay plot in contrasting color to the scatter plot, the first coordinate equal to the first utility measure computed on the pairs of documents including the additional document the second coordinate is equal to the second utility measure computed on the pairs of documents including the additional document.

36. The method of claim 35 wherein said additional document is a textual query entered by a user.

37. The method of claim 35 wherein:

the first representation is a conceptual-level representation; and

the second representation is a term-based representation.

38. The method of claim 37 wherein:

the first representation is a subject vector; and

the second representation is a word vector.

39. The method of claim 19 wherein said step of determining a first utility measure further comprises the steps of:

determining a first intermediate utility measure;

determining a second intermediate utility measure;

selecting a particular intermediate utility measure from said first intermediate utility measure and said second intermediate utility measure as said first utility measure.

40. The method of claim 29 wherein said step of determining a second utility measure further comprises the following steps:

determining a third intermediate utility measure;

determining a fourth intermediate utility measure;

selecting a particular intermediate utility measure from said third intermediate utility measure and said fourth intermediate utility measure as said second utility measure.

41. The method of claim 39 wherein:

said first intermediate utility measure is a combination of a first similarity measure for said first document element and said first similarity measure for said second document element and a first normalization constant; and

said second intermediate utility measure is a combination of a first similarity measure for said second document element and said first similarity measure for said first document element and a second normalization constant.

42. The method of claim 40 wherein:

(a) said third intermediate utility measure is a combination of a second similarity measure for said first document element and said second similarity measure for said second document element and a first normalization constant; and

(b) said fourth intermediate utility measure is a combination of said second similarity measure for said second document element and said second similarity measure for said first document element and a second normalization constant.

43. The method of claim 41 wherein said first similarity measure is a word weight vector.

44. The method of claim 42 wherein said second similarity measure is an SFC weight vector.

45. The method of claim 19 wherein:

said pairs of documents further comprises a first document and a second document,

said first document is a dependent claim, x, depending from an independant claim, X, and

said second document is a dependent claim, y, depending from an independent claim, Y,

said determining a first utility measure further comprises the following steps:

determining a first intermediate utility measure;

determining a second intermediate utility measure;

combining said first intermediate utility measure and said second intermediate utility measure.

46. The method of claim 29 wherein:

said pairs of documents further comprises a first document and a second document,

said first document is a dependent claim, x, depending from an independant claim, X, and

said second document is a dependent claim, y, depending from an independent claim, Y,

said determining a second utility measure further comprises the following steps:

determining a third intermediate utility measure;

determining a fourth intermediate utility measure;

combining said third intermediate utility measure and said fourth intermediate utility measure.

47. The method of claim 45 wherein:

(a) said first intermediate utility measure is a combination of a first similarity measure for said first document element, x, and said first similarity measure for said second document element Y; and

(b) said second intermediate utility measure is a combination of said first similarity measure for said first document element, X, and said first similarity measure for said second document element, Y.

48. The method of claim 46 wherein:

(a) said third intermediate utility measure is a combination of said second similarity measure for said first document element, x, and said second similarity measure for said second document element Y; and

(b) said fourth intermediate utility measure is a combination of said second similarity measure for said first document element, X, and said second similarity measure for said second document element, Y.

49. The method of claim 47 wherein said first similarity measure is a word weight vector.

50. The method of claim 48 wherein said second similarity measure an SFC weight vector.

51. The method of claim 45 wherein said step of combining comprises an averaging.

52. The method of claim 46 wherein said step of combining comprises an averaging.

53. A computer program product which analyzes and displays information regarding a plurality of documents comprising:

code for generating first and second representations of each document;

code for determining, for selected pairs of documents;

(a) a first utility score based on the respective first representations of the documents in that pair, and

(b) a second utility score based on the respective second representations of the documents in that pair;

code for displaying a scatter plot in an area bounded by a first and a second non-parallel axes where each selected pair is represented by a point having a first coordinate along the first axis equal to the first utility score and a second coordinate along the second axis equal to the second utility score; and

a computer readable storage medium for storing the codes.

54. A method of analyzing patent documents comprising the steps of:

providing a dataset containing a plurality of patent documents;

identifying within each patent document a portion of said document containing a set of claims;

generating a first representation of each set of claims within said plurality of patent documents; and

determining a first utility measure of at least one claim within at least one set of claims based upon similarity of said at least one claim with a query document.

55. The method of claim 54 wherein said query document is a concept query, patent or claim.

56. The method of claim 54 further comprising the step of displaying on a computer screen a ranking of a plurality of claims contained within said patent documents based upon said first utility measure associated with each of said plurality of claims, said screen including a claim number and rank number for each of said plurality of claims.

57. The method of claim 56 further comprising the step of providing a link at said claim number to a full-text display of an associated claim.

58. The method of claim 57 further comprising the step of providing a link at said rank number to a side-by-side textual display of said associated claim and said query document.

59. The method of claim 54 further comprising the step of parsing each set of claims to identify each individual claim within said each set and all claims referenced by said each individual claim.

60. The method of claim 54 further comprising the steps of:

generating a second representation of each set of claims within said plurality of patent documents; and

determining a second utility measure of said at least one claim within said at least one set of claims based upon similarity of said at least one claim with said query document.

61. The method of claim 60 further comprising the step of displaying on a computer screen a ranking of said plurality of patent documents based upon said first and second utility measures associated with claims of each of said patent documents, said screen including a rank number for each of said plurality of patent documents.

62. The method of claim 61 further comprising the step of providing a link at said rank number to a side-by-side textual display of an associated patent document and said query document.

63. The method of claim 62 further comprising the step of providing a link at a screen icon to a textual display of a ranked listing of matching claims of said associated patent document and said query document.

64. A method of analyzing a patent document comprising the steps of:

providing a dataset containing at least one patent document;

identifying within said at least one patent document a portion of said document containing a set of claims;

parsing said set of claims to identify an individual claim within said set and all claims referenced by said individual claim; and

displaying on a computer screen a link for each claim referenced by said individual claim.

65. The method of claim 64 further comprising the step of displaying on said computer screen at least a portion of said individual claim.

66. The method of claim 65 wherein activation of said link for a particular claim referenced by said individual claim produces a full text display of said particular claim.

67. The method of claim 66 wherein said link is a claim number.

68. The method of claim 67 wherein said full text display of said particular claim comprises a transitive closure of said particular claim.

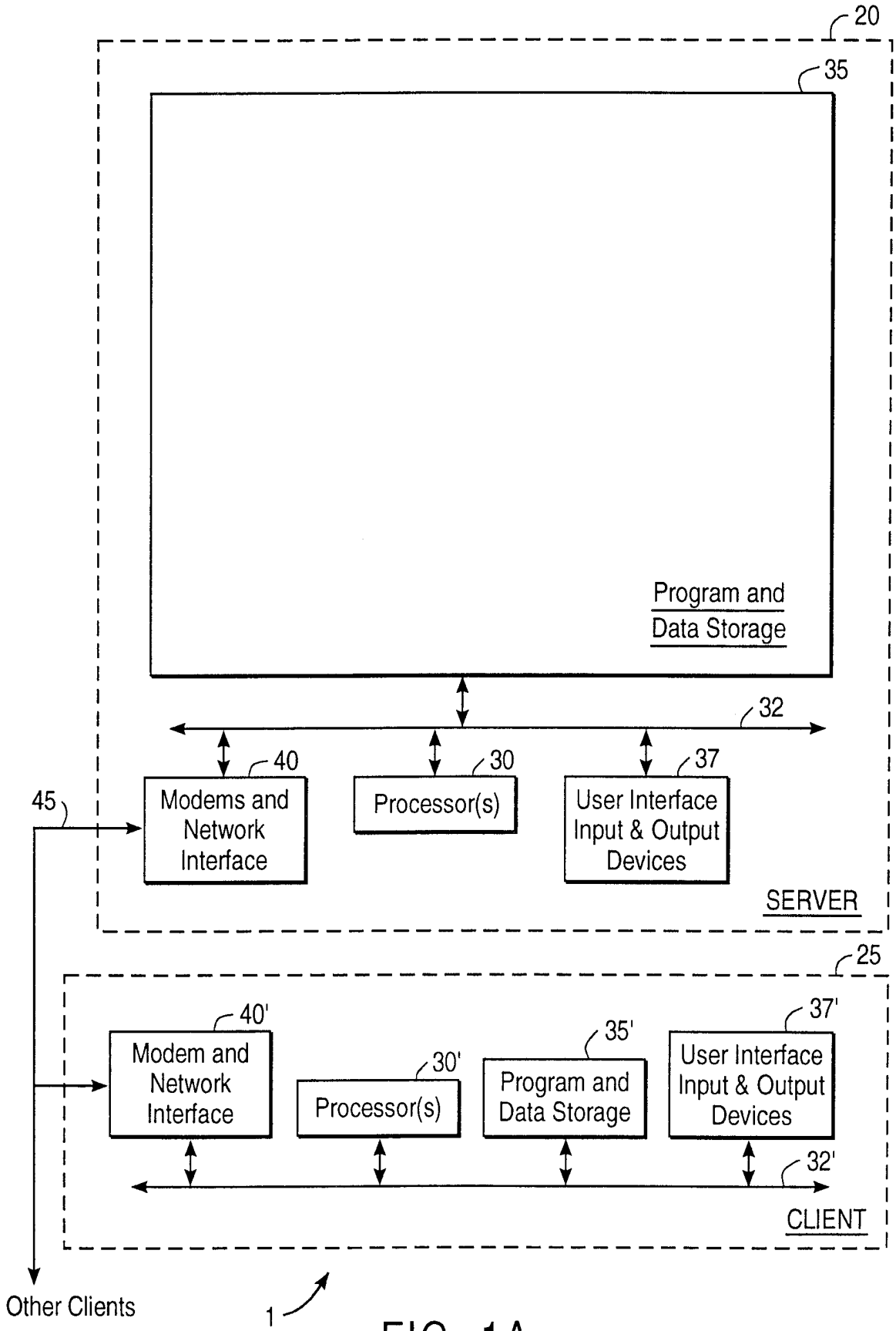

20
35
Program and
Data Storage
32
40
30
37
45
Modems and
Network
Interface
Processor(s)
User Interface
Input & Output
Devices
SERVER
25
40'
37'
35'
Modem and
Network
Interface
30'
Processor(s)
Program and
Data Storage
User Interface
Input & Output
Devices
32'
CLIENT
Other Clients
1

FIG. 1A

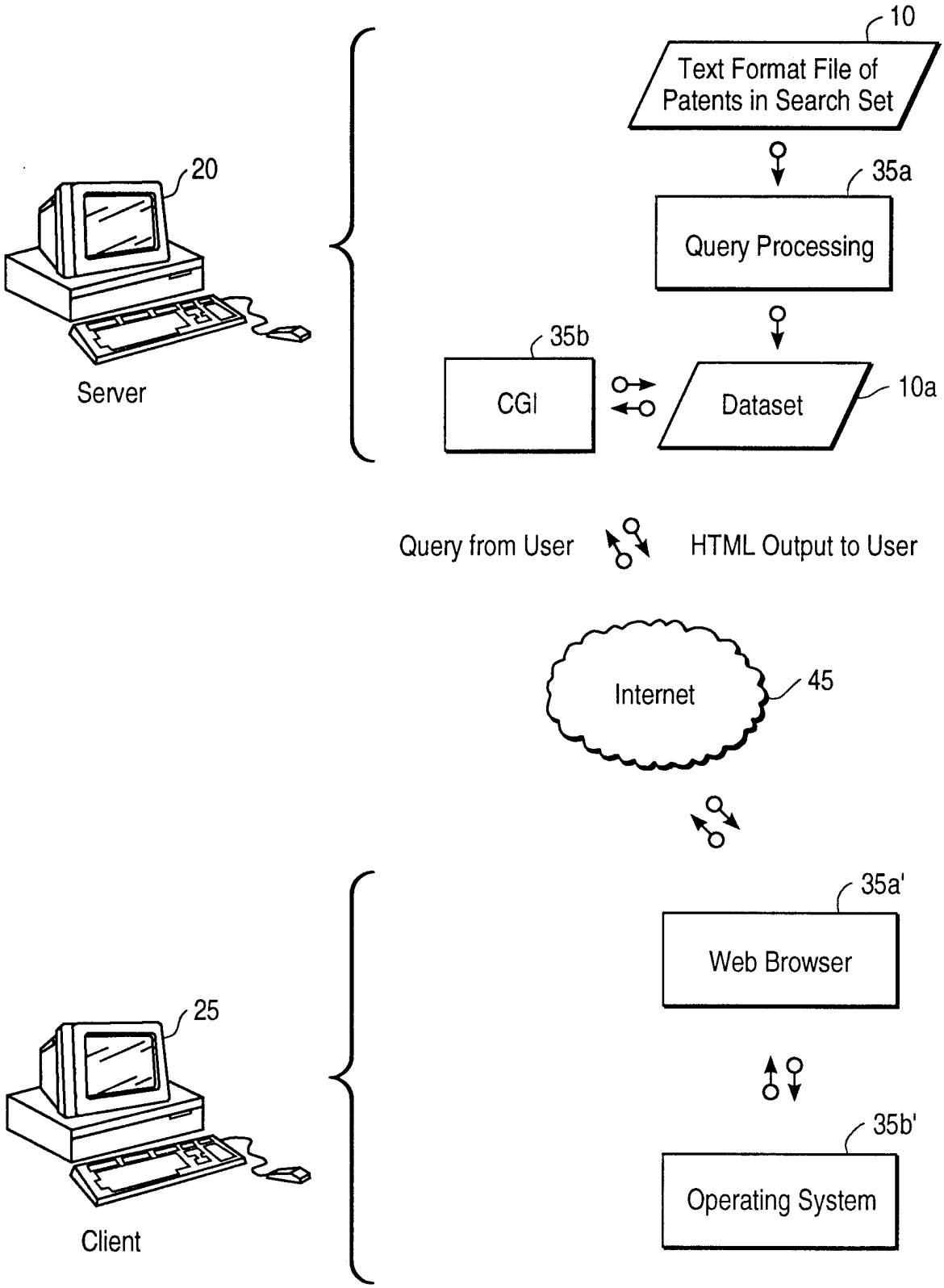

10
Text Format File of Patents in Search Set
35a
Query Processing
35b
CGI
Dataset
10a
20
Server
Query from User
HTML Output to User
Internet
45
35a'
Web Browser
35b'
Operating System
25
Client

FIG. 1B

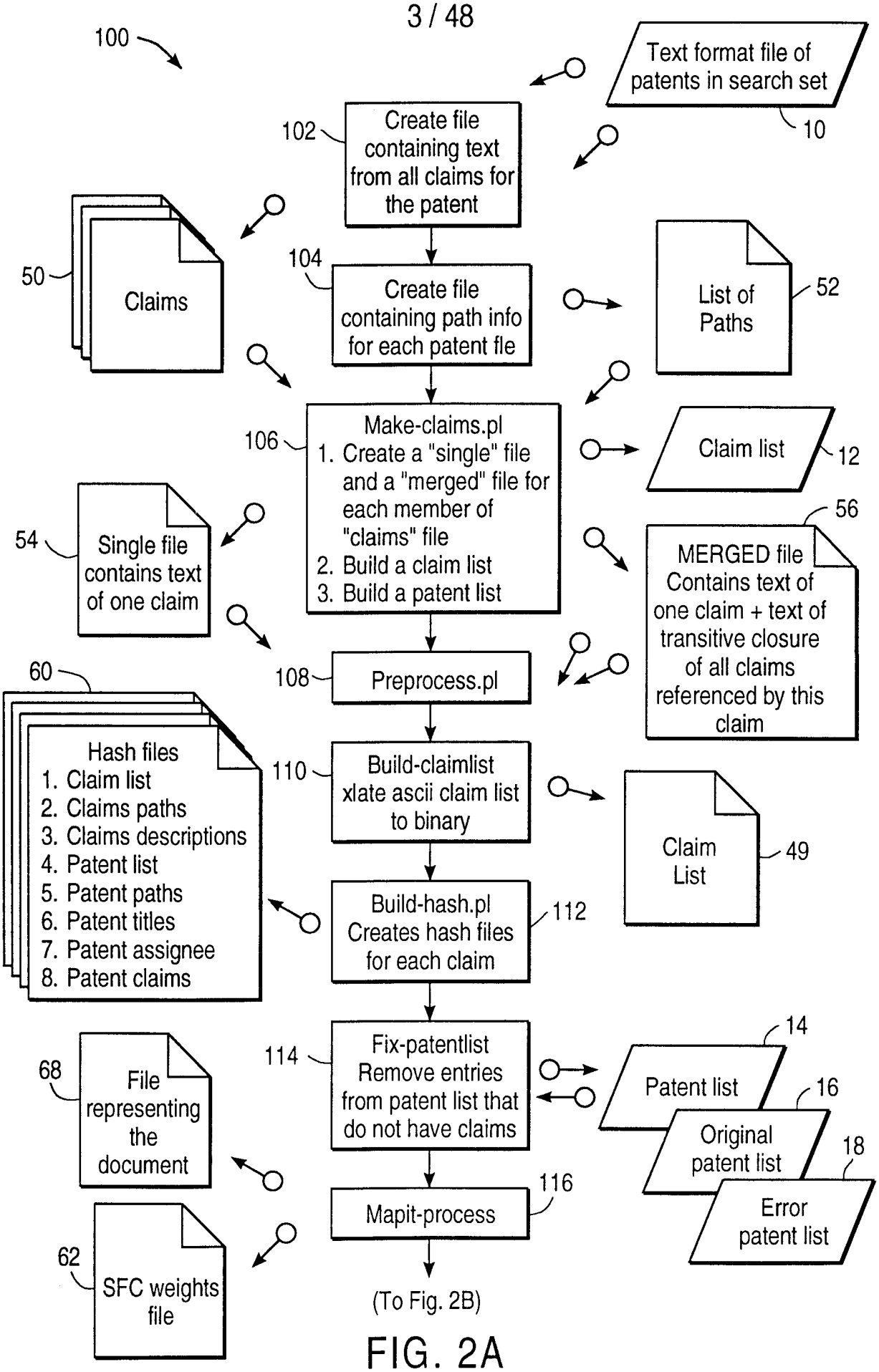

100
Text format file of patents in search set
10
102
Create file containing text from all claims for the patent
50
Claims
104
Create file containing path info for each patent fle
List of Paths
52
106
Make-claims.pl
1. Create a "single" file and a "merged" file for each member of "claims" file
2. Build a claim list
3. Build a patent list
Claim list
12
54
Single file contains text of one claim
56
MERGED file Contains text of one claim + text of transitive closure of all claims referenced by this claim
108
Preprocess.pl
60
Hash files
1. Claim list
2. Claims paths
3. Claims descriptions
4. Patent list
5. Patent paths
6. Patent titles
7. Patent assignee
8. Patent claims
110
Build-claimlist xlate ascii claim list to binary
Claim List
49
Build-hash.pl Creates hash files for each claim
112
14
114
Fix-patentlist Remove entries from patent list that do not have claims
Patent list
16
Original patent list
18
Error patent list
68
File representing the document
116
Mapit-process
62
SFC weights file
(To Fig. 2B)

FIG. 2A

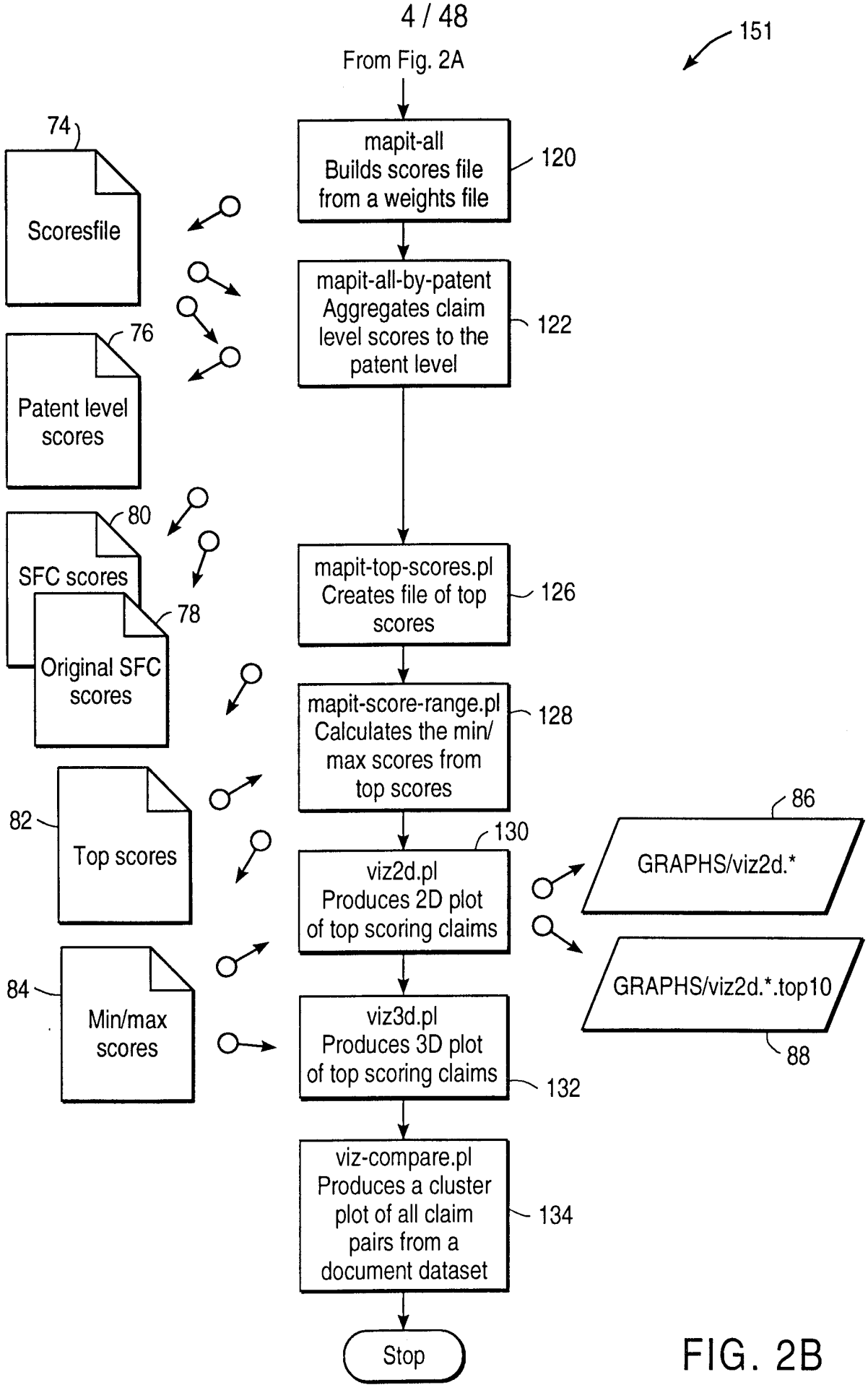

151
From Fig. 2A
mapit-all
Builds scores file from a weights file
120
mapit-all-by-patent
Aggregates claim level scores to the patent level
122
mapit-top-scores.pl
Creates file of top scores
126
mapit-score-range.pl
Calculates the min/ max scores from top scores
128
viz2d.pl
Produces 2D plot of top scoring claims
130
viz3d.pl
Produces 3D plot of top scoring claims
132
viz-compare.pl
Produces a cluster plot of all claim pairs from a document dataset
134
Stop
74
Scoresfile
76
Patent level scores
80
SFC scores
78
Original SFC scores
82
Top scores
84
Min/max scores
86
GRAPHS/viz2d.*
GRAPHS/viz2d.*.top10
88

FIG. 2B

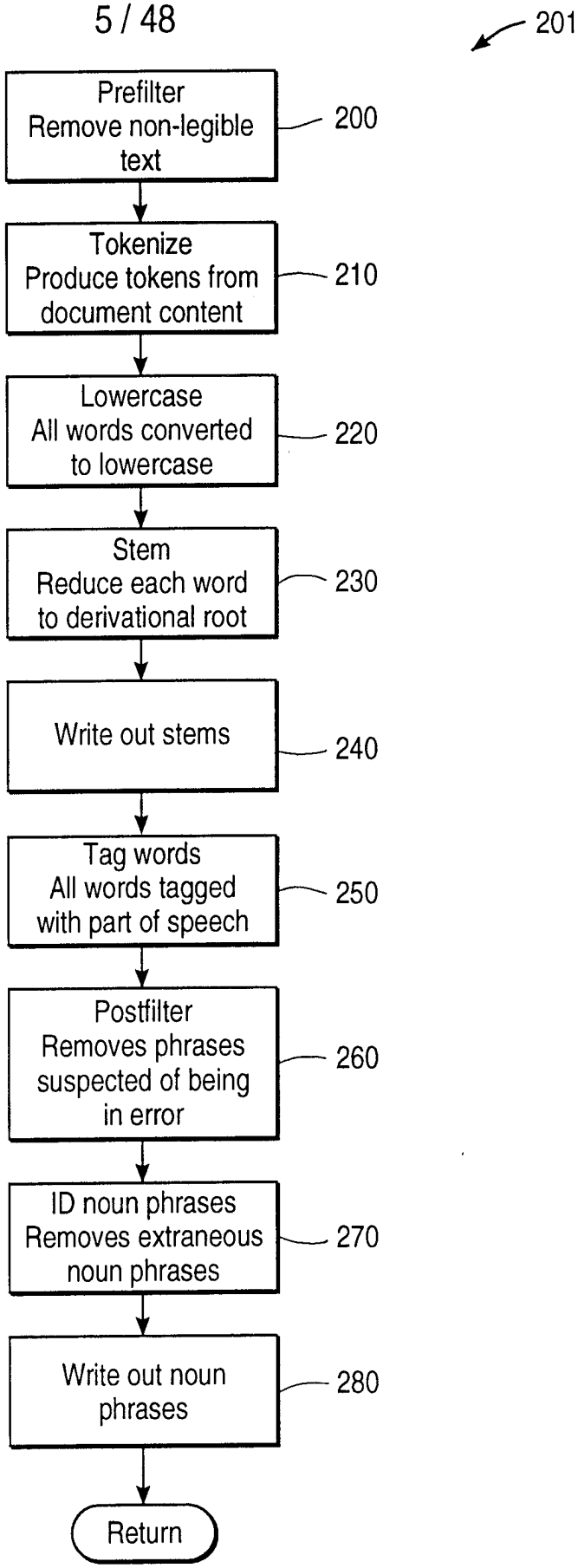

201
Prefilter
Remove non-legible text
200
Tokenize
Produce tokens from document content
210
Lowercase
All words converted to lowercase
220
Stem
Reduce each word to derivational root
230
Write out stems
240
Tag words
All words tagged with part of speech
250
Postfilter
Removes phrases suspected of being in error
260
ID noun phrases
Removes extraneous noun phrases
270
Write out noun phrases
280
Return

FIG. 3

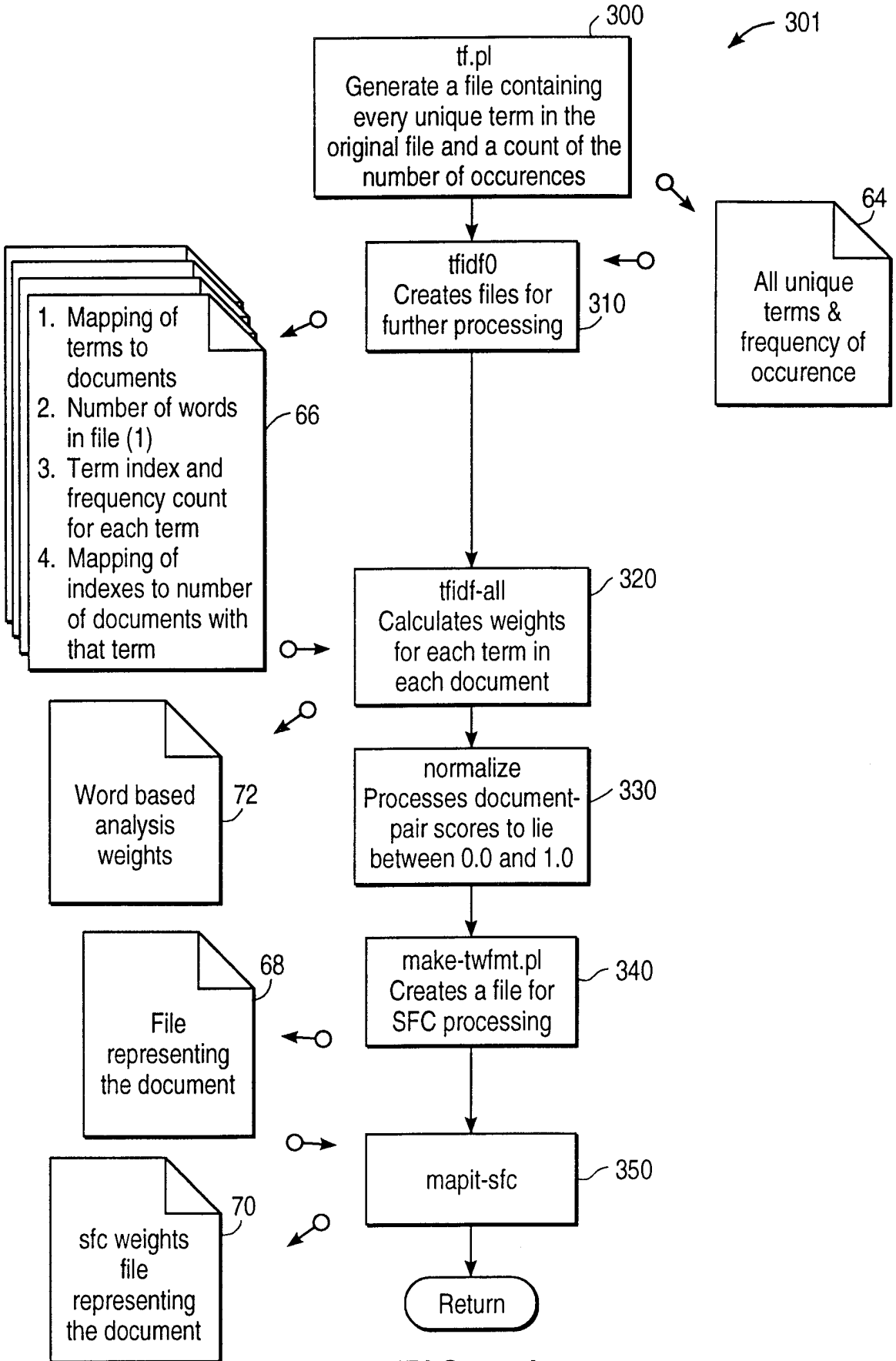

300
301
tf.pl
Generate a file containing every unique term in the original file and a count of the number of occurences
64
All unique terms & frequency of occurence
tfidf0
Creates files for further processing
310
1. Mapping of terms to documents
2. Number of words in file (1)
3. Term index and frequency count for each term
4. Mapping of indexes to number of documents with that term
66
tfidf-all
Calculates weights for each term in each document
320
Word based analysis weights
72
normalize
Processes document-pair scores to lie between 0.0 and 1.0
330
make-twfmt.pl
Creates a file for SFC processing
340
68
File representing the document
mapit-sfc
350
70
sfc weights file representing the document
Return

FIG. 4A

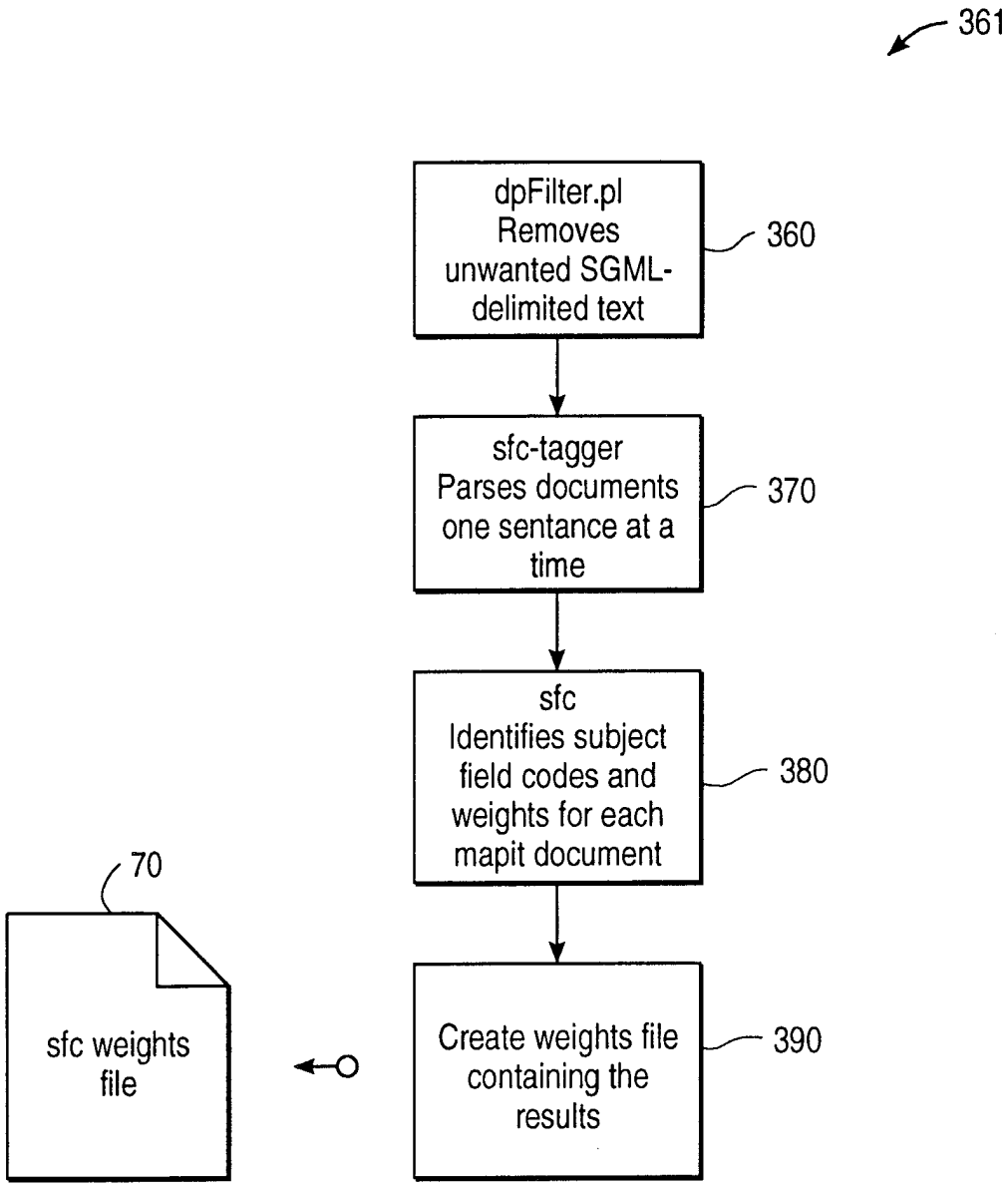

361
dpFilter.pl
Removes
unwanted SGML-
delimited text
360
sfc-tagger
Parses documents
one sentance at a
time
370
sfc
Identifies subject
field codes and
weights for each
mapit document
380
70
sfc weights
file
Create weights file
containing the
results
390

FIG. 4B

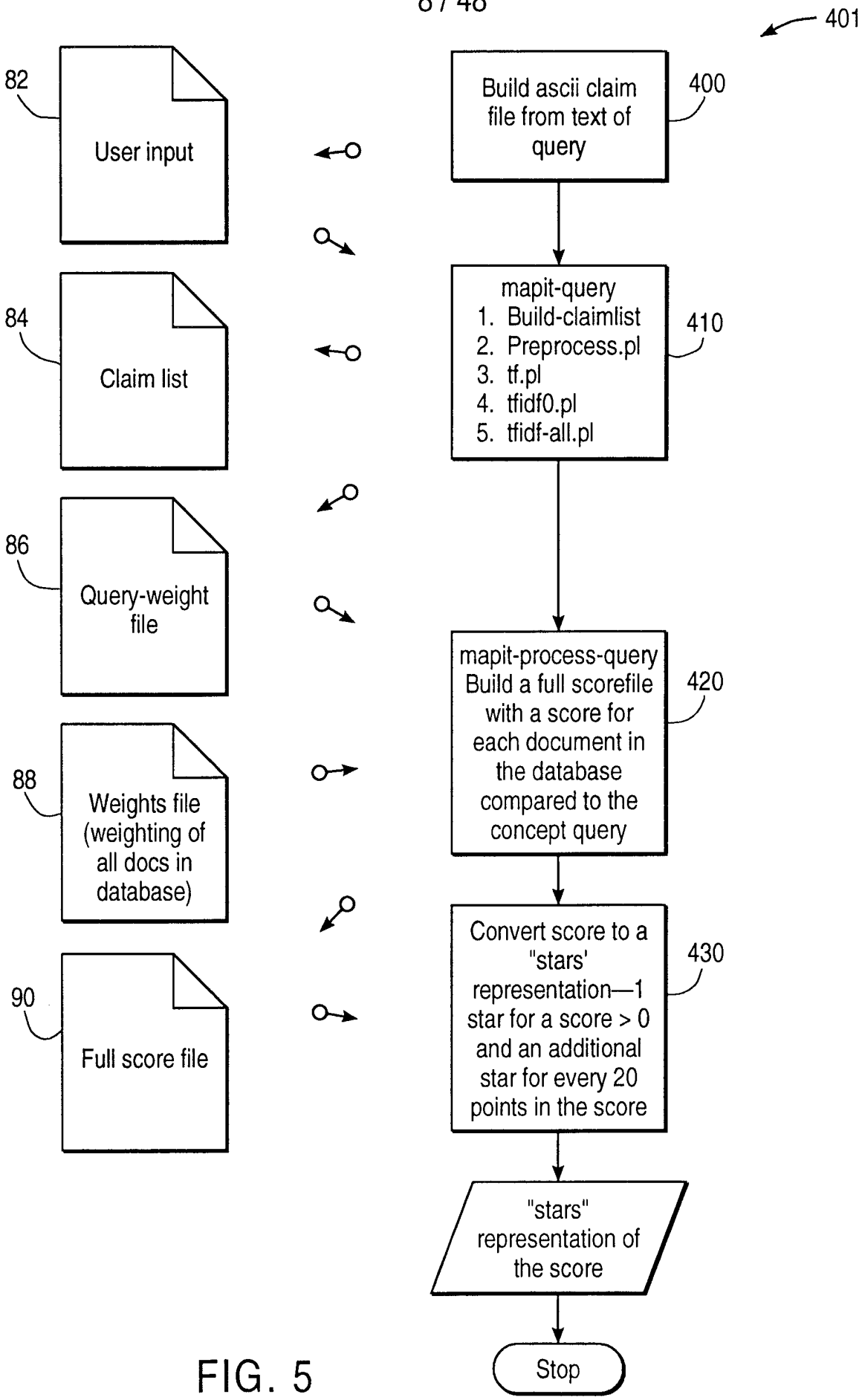

401
82
User input
84
Claim list
86
Query-weight file
88
Weights file (weighting of all docs in database)
90
Full score file
Build ascii claim file from text of query
400
mapit-query
1. Build-claimlist
2. Preprocess.pl
3. tf.pl
4. tfidf0.pl
5. tfidf-all.pl
410
mapit-process-query Build a full scorefile with a score for each document in the database compared to the concept query
420
Convert score to a "stars' representation—1 star for a score > 0 and an additional star for every 20 points in the score
430
"stars" representation of the score
Stop

FIG. 5

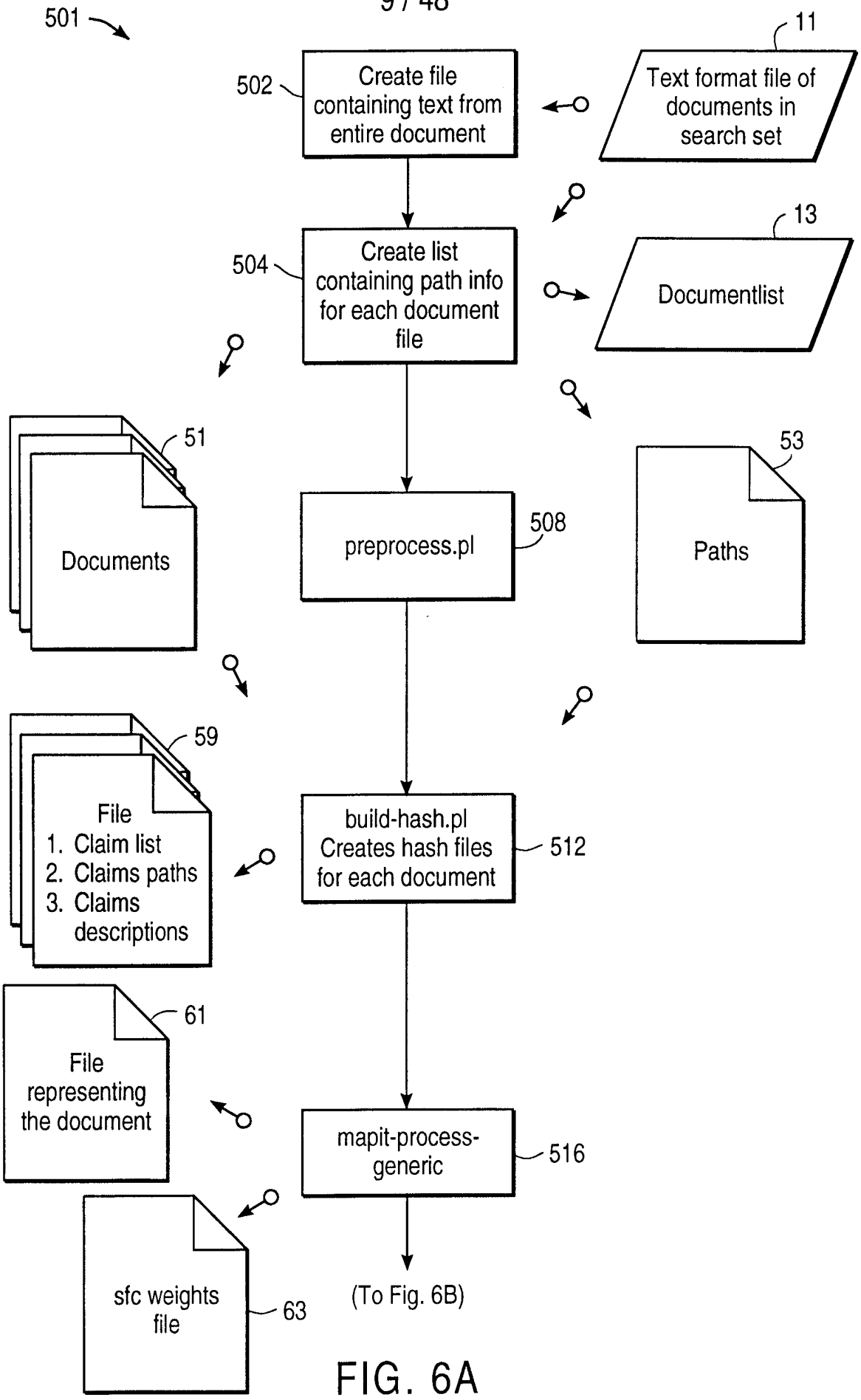

501
502
Create file containing text from entire document
11
Text format file of documents in search set
504
Create list containing path info for each document file
13
Documentlist
51
Documents
53
Paths
preprocess.pl
508
59
File
1. Claim list
2. Claims paths
3. Claims descriptions
build-hash.pl
Creates hash files for each document
512
61
File representing the document
mapit-process-generic
516
sfc weights file
63
(To Fig. 6B)

FIG. 6A

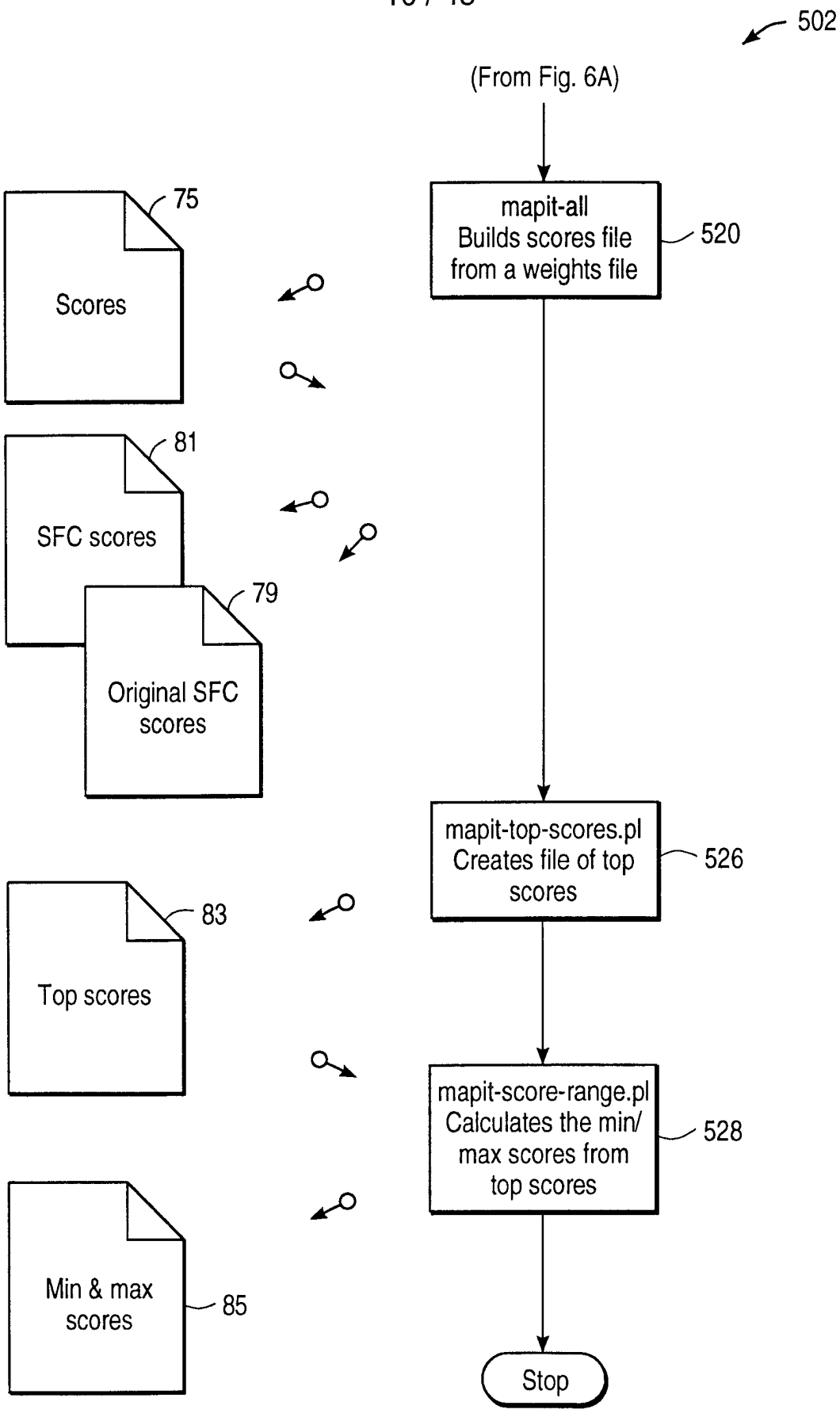

502
(From Fig. 6A)
mapit-all
Builds scores file
from a weights file
520
75
Scores
81
SFC scores
79
Original SFC
scores
mapit-top-scores.pl
Creates file of top
scores
526
83
Top scores
mapit-score-range.pl
Calculates the min/
max scores from
top scores
528
Min & max
scores
85
Stop

FIG. 6B

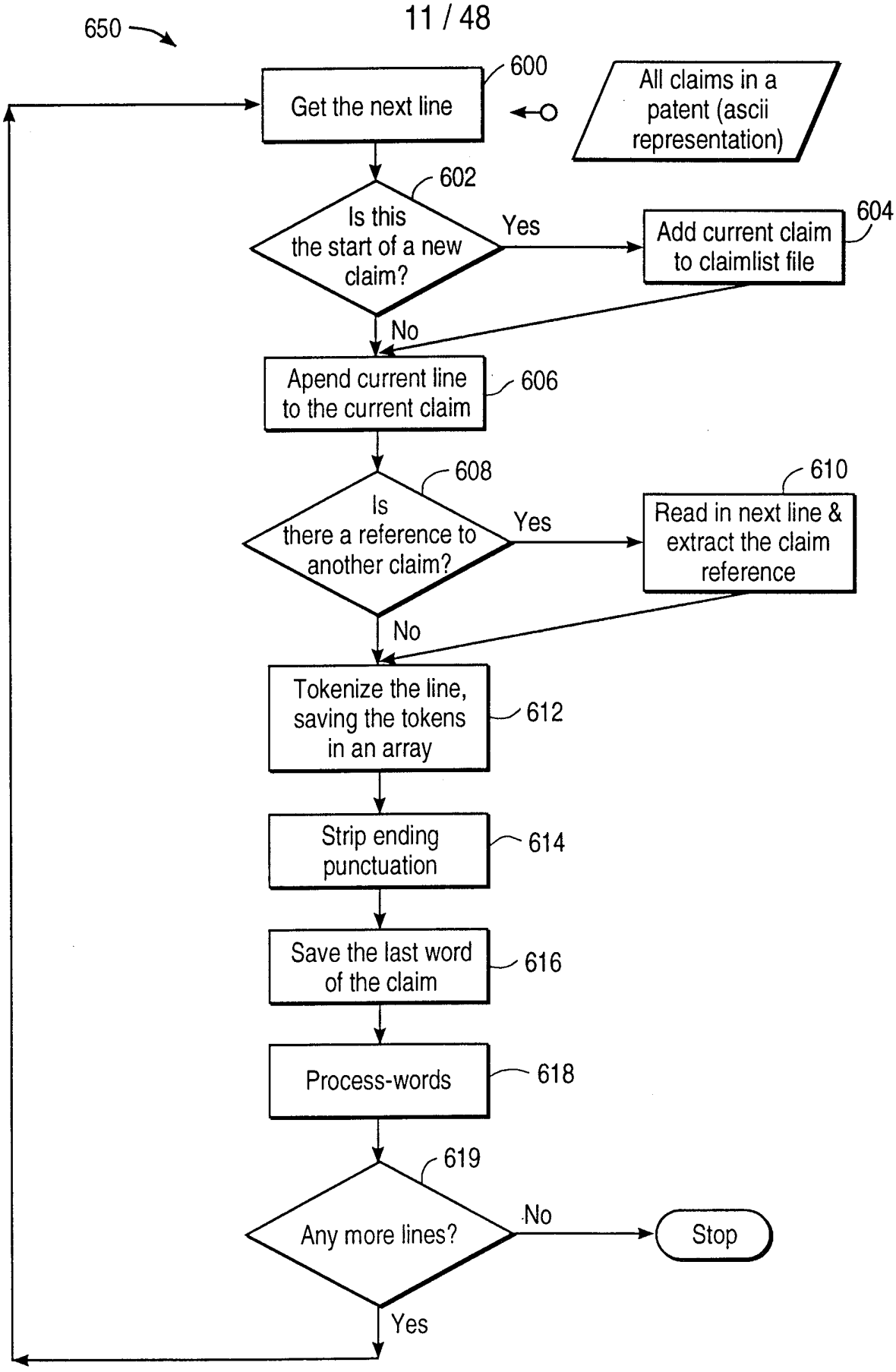

650
600
Get the next line
All claims in a patent (ascii representation)
602
Is this the start of a new claim?
Yes
604
Add current claim to claimlist file
No
Apend current line to the current claim
606
608
Is there a reference to another claim?
Yes
610
Read in next line & extract the claim reference
No
Tokenize the line, saving the tokens in an array
612
Strip ending punctuation
614
Save the last word of the claim
616
Process-words
618
619
Any more lines?
No
Stop
Yes

FIG. 7A

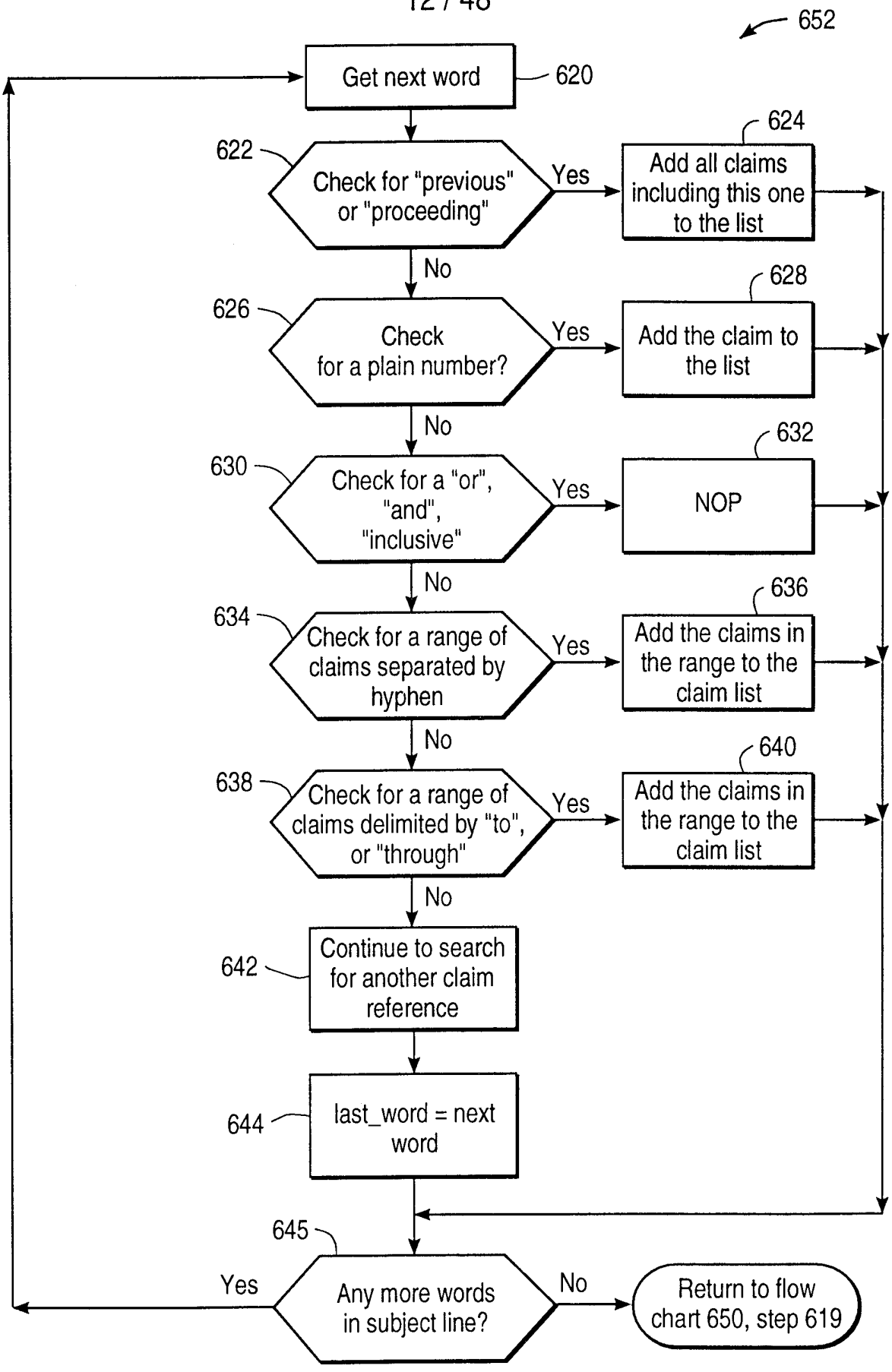

652
Get next word
620
622
Check for "previous" or "proceeding"
Yes
624
Add all claims including this one to the list
No
626
Check for a plain number?
Yes
628
Add the claim to the list
No
630
Check for a "or", "and", "inclusive"
Yes
632
NOP
No
634
Check for a range of claims separated by hyphen
Yes
636
Add the claims in the range to the claim list
No
638
Check for a range of claims delimited by "to", or "through"
Yes
640
Add the claims in the range to the claim list
No
642
Continue to search for another claim reference
644
last_word = next word
645
Yes
Any more words in subject line?
No
Return to flow chart 650, step 619

FIG. 7B

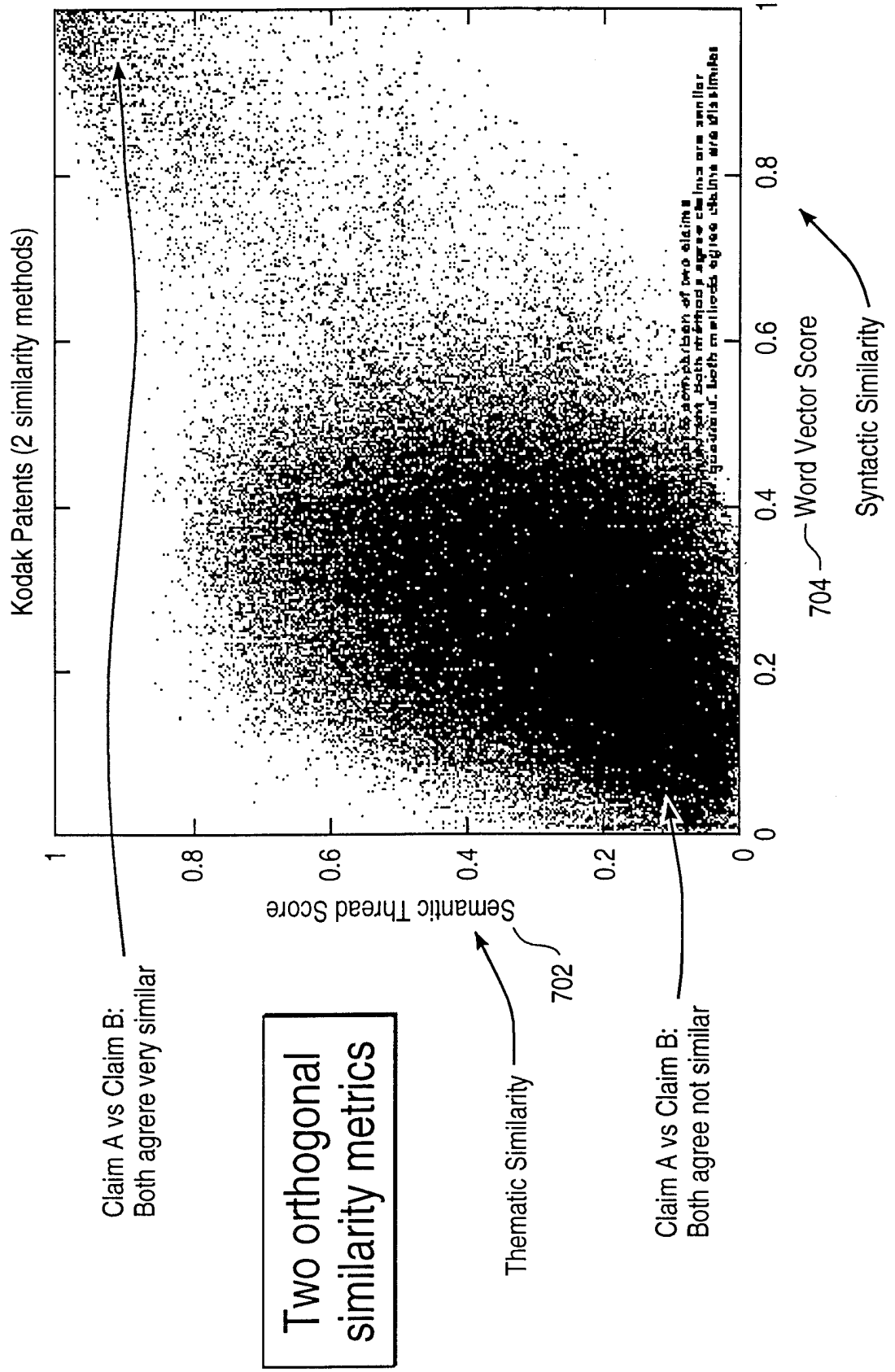

Kodak Patents (2 similarity methods)
Semantic Thread Score
Word Vector Score
Syntactic Similarity
Claim A vs Claim B:
Both agrere very similar
Claim A vs Claim B:
Both agree not similar
Thematic Similarity
702
704
Two orthogonal
similarity metrics
0
0.2
0.4
0.6
0.8
1

FIG. 8A

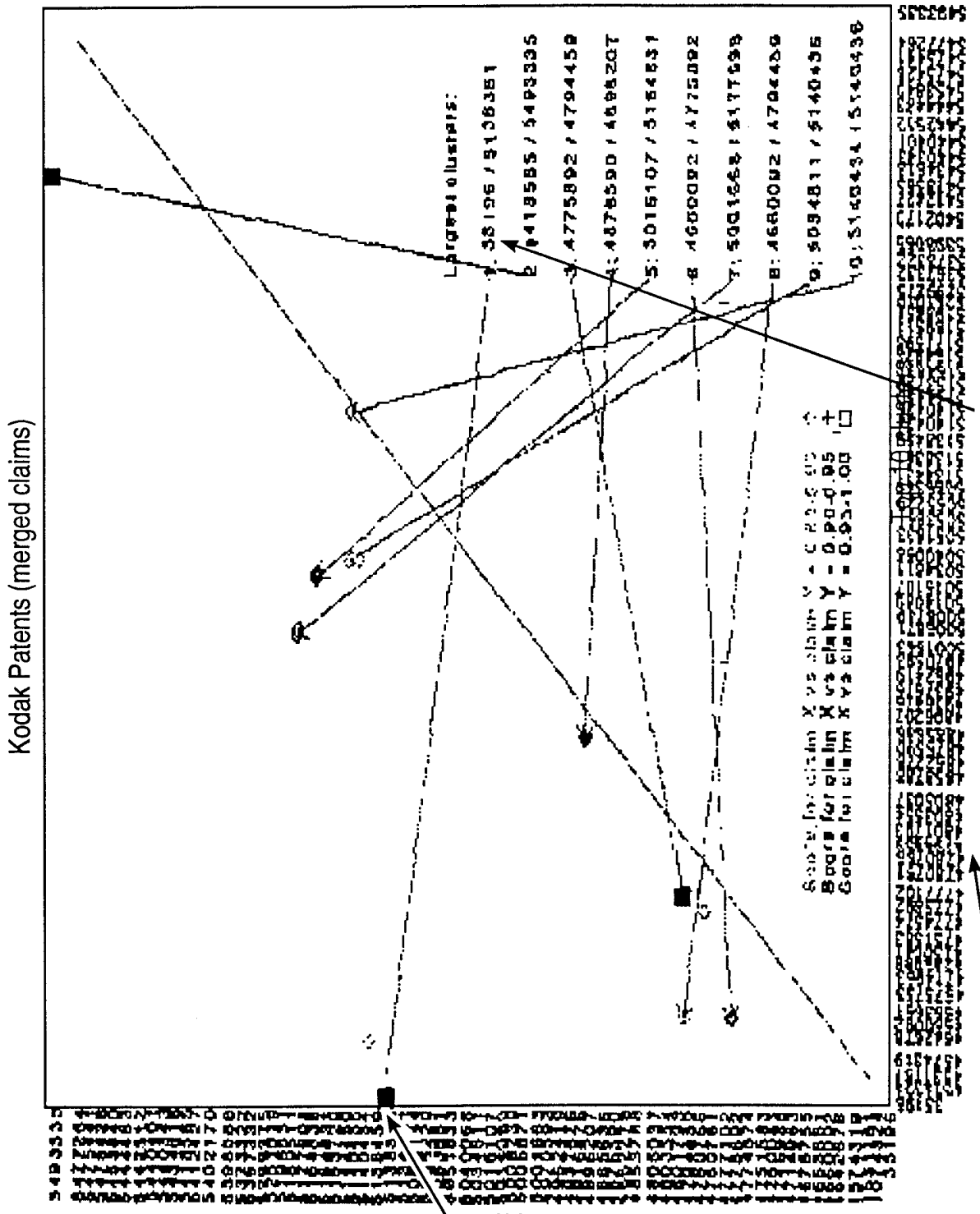

2D clustering of similar patents
Kodak Patents (merged claims)
Patent Numbers
Cluster of similar claims (0.95)
Patent set A
Patent set B
Top 10 clusters

FIG. 8B

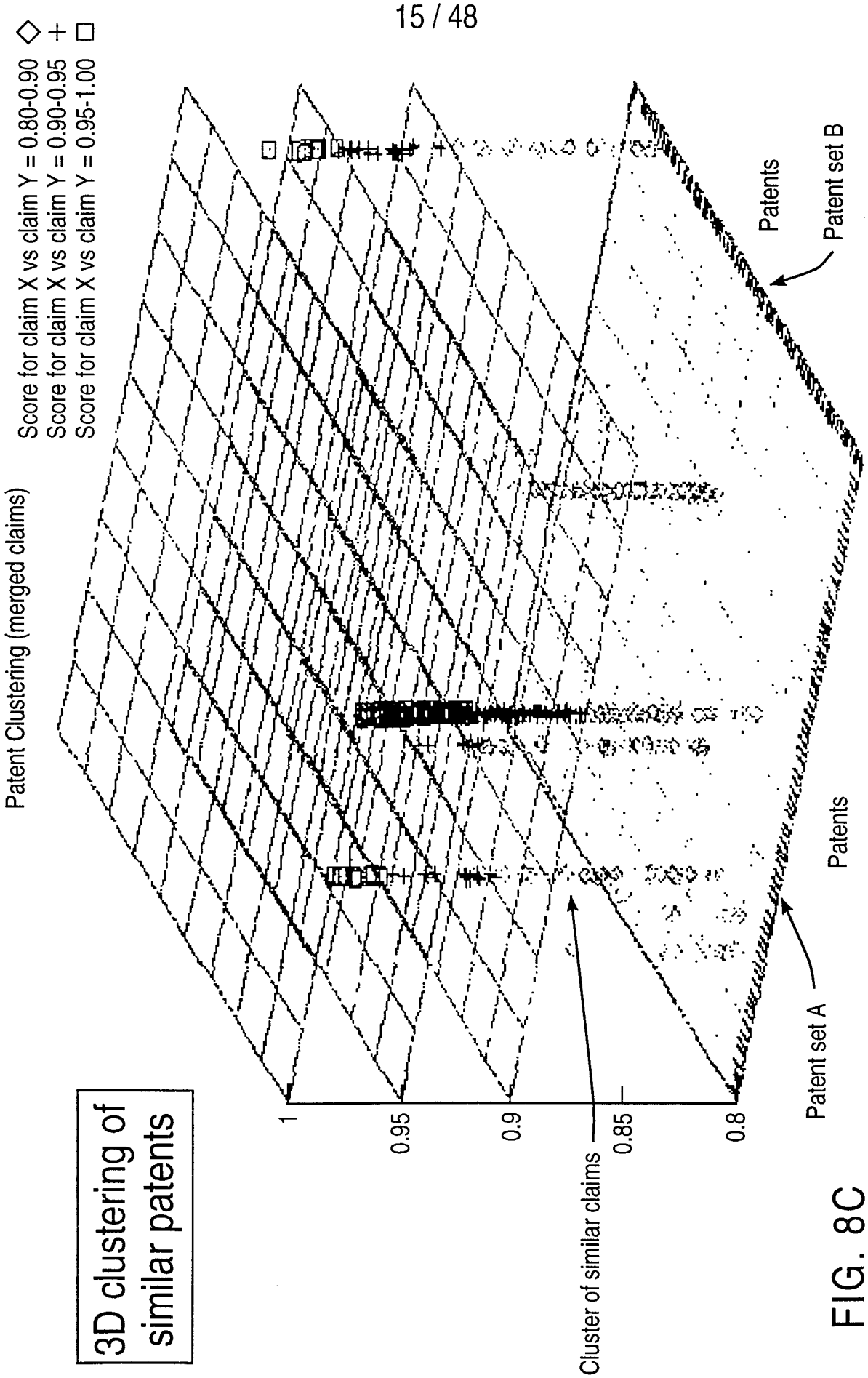

3D clustering of similar patents
Patent Clustering (merged claims)
Score for claim X vs claim Y = 0.80-0.90
Score for claim X vs claim Y = 0.90-0.95
Score for claim X vs claim Y = 0.95-1.00
1
0.95
0.9
0.85
0.8
Cluster of similar claims
Patent set A
Patents
Patent set B
Patents

FIG. 8C

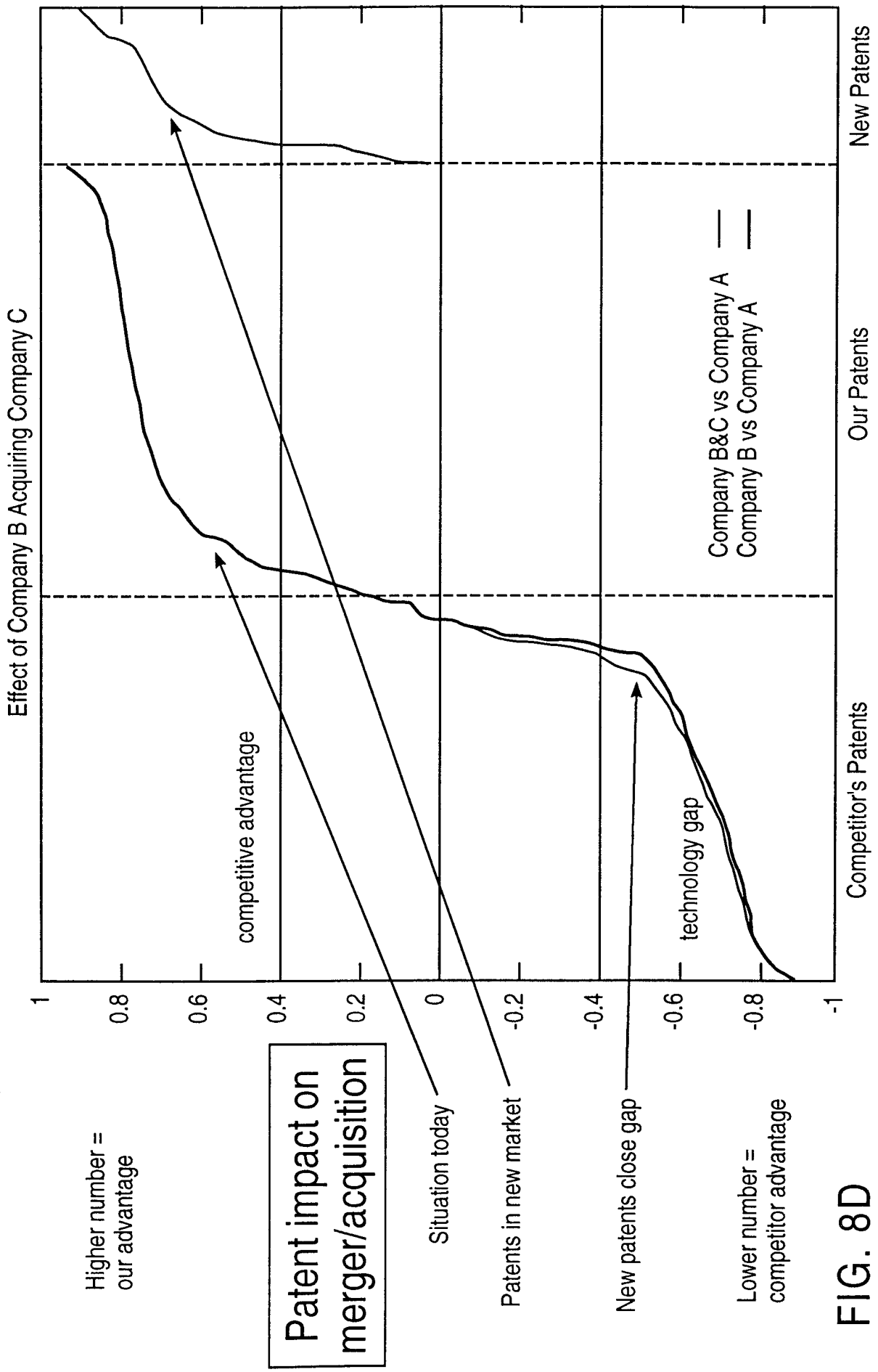

Effect of Company B Acquiring Company C
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
competitive advantage
technology gap
Company B&C vs Company A
Company B vs Company A
Our Patents
New Patents
Competitor's Patents
Patent impact on merger/acquisition
Higher number = our advantage
Lower number = competitor advantage
Situation today
Patents in new market
New patents close gap

FIG. 8D

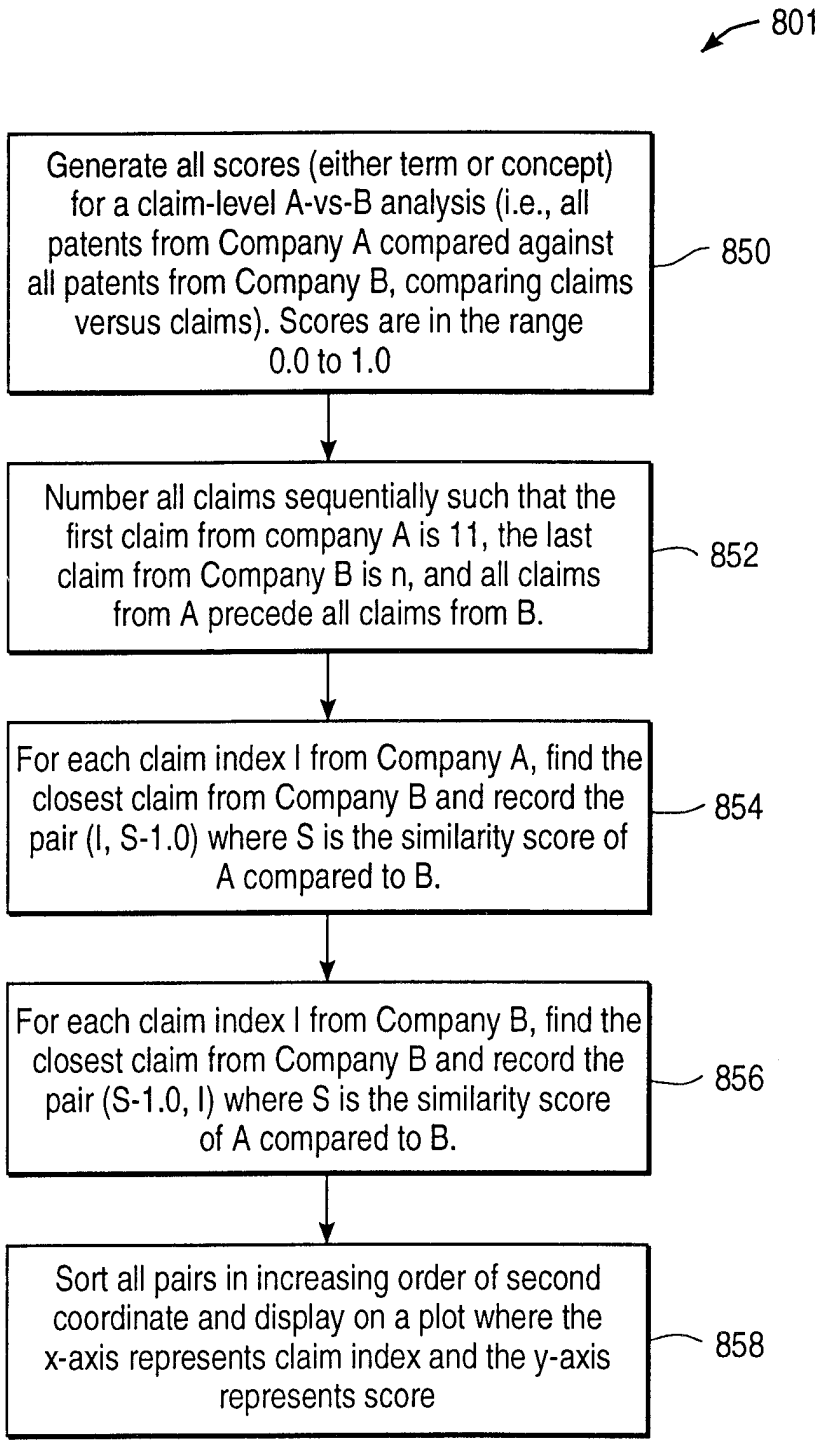
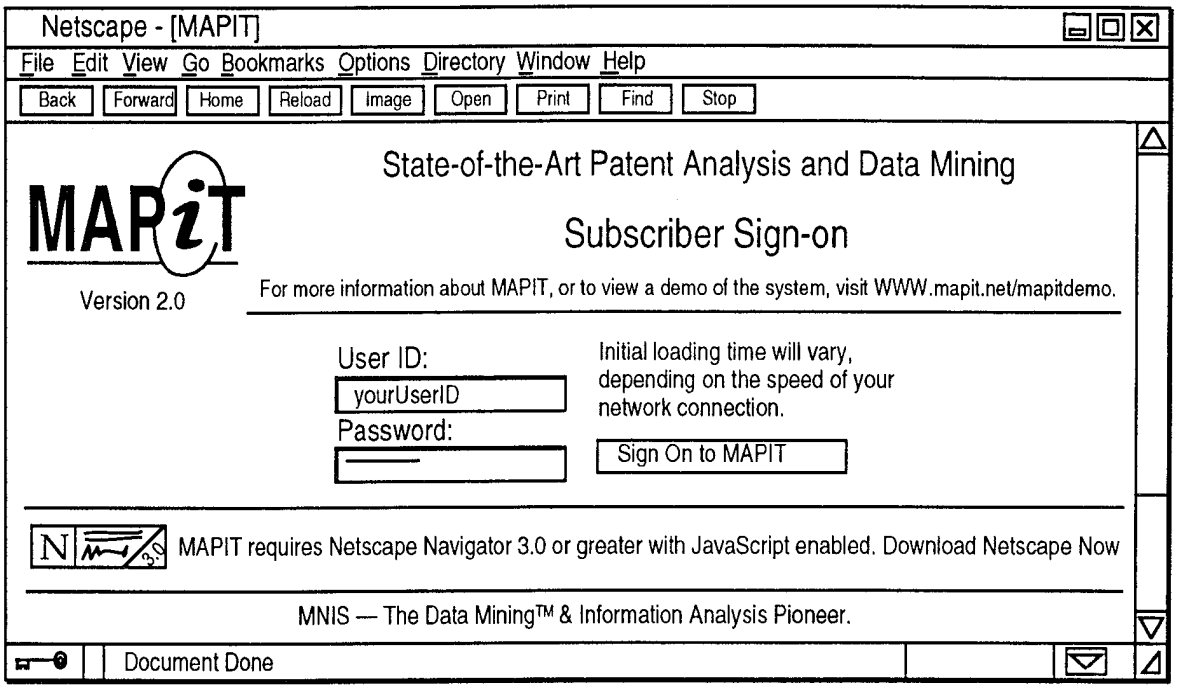

801
Generate all scores (either term or concept) for a claim-level A-vs-B analysis (i.e., all patents from Company A compared against all patents from Company B, comparing claims versus claims). Scores are in the range 0.0 to 1.0
850
Number all claims sequentially such that the first claim from company A is 11, the last claim from Company B is n, and all claims from A precede all claims from B.
852
For each claim index I from Company A, find the closest claim from Company B and record the pair (I, S-1.0) where S is the similarity score of A compared to B.
854
For each claim index I from Company B, find the closest claim from Company B and record the pair (S-1.0, I) where S is the similarity score of A compared to B.
856
Sort all pairs in increasing order of second coordinate and display on a plot where the x-axis represents claim index and the y-axis represents score
858

FIG. 8E



Signing On

Figure 1.1 The MAPIT Sign-on Screen:
Showing the Netscape Secure Server version
of the system.

FIG. 9A

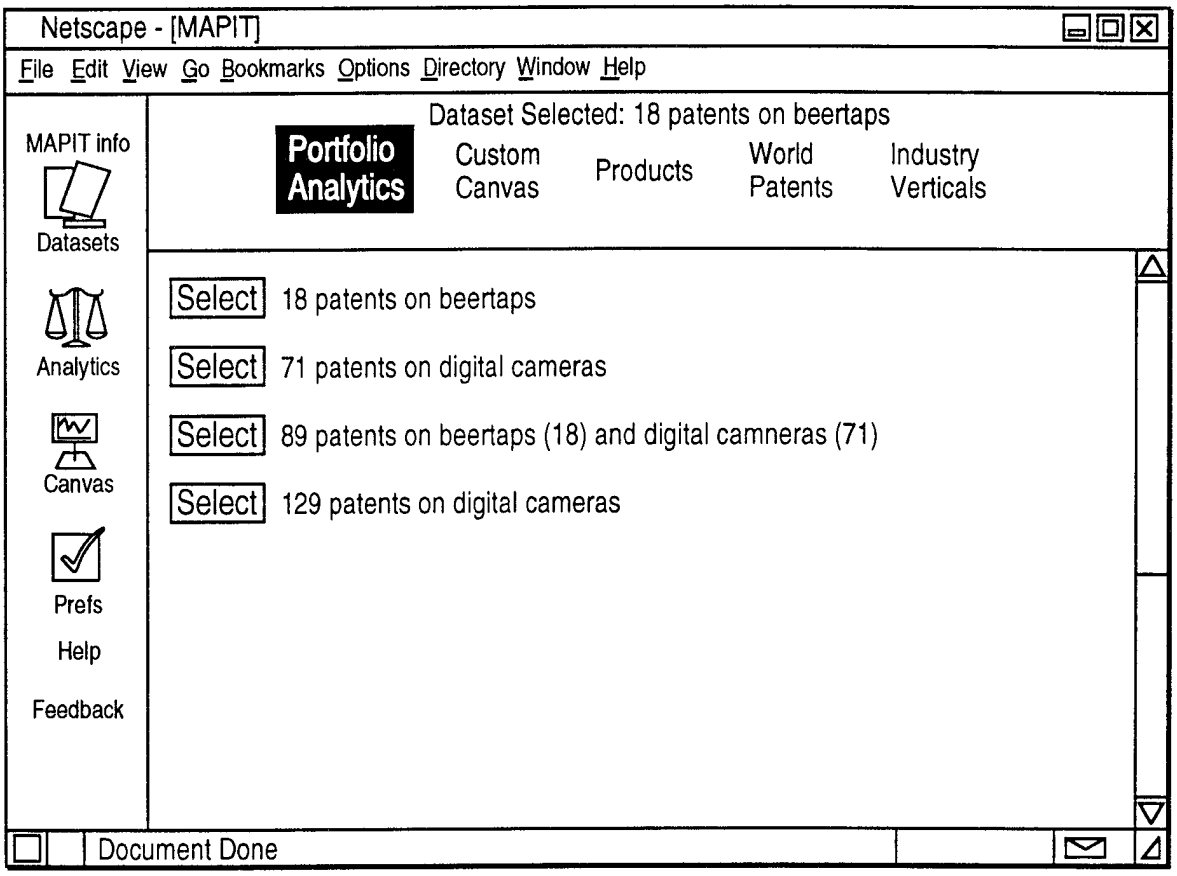

Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Dataset Selected: 18 patents on beertaps
Portfolio Analytics
Custom Canvas
Products
World Patents
Industry Verticals
Select 18 patents on beertaps
Select 71 patents on digital cameras
Select 89 patents on beertaps (18) and digital camneras (71)
Select 129 patents on digital cameras
Document Done

FIG. 9B

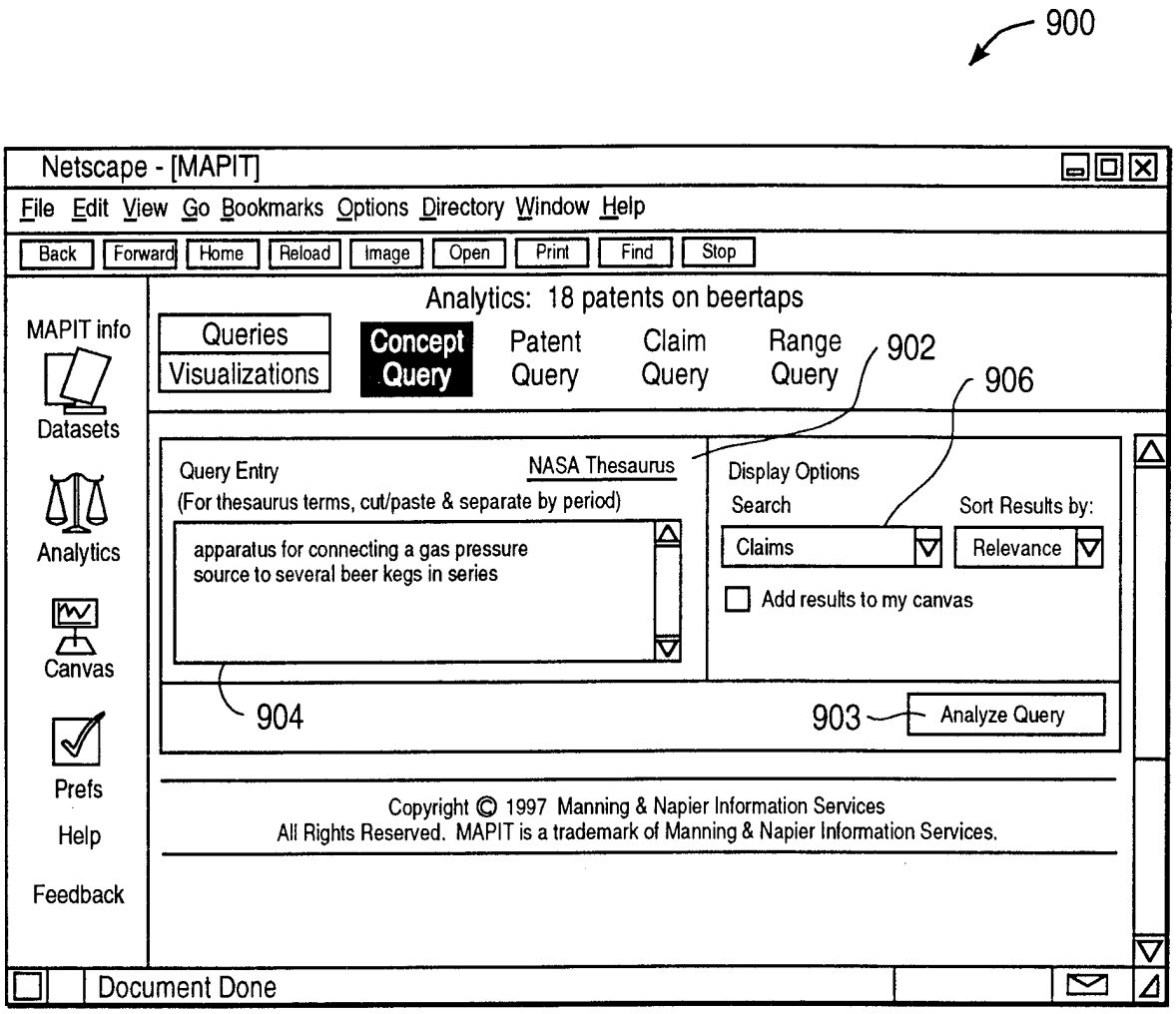

900
Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
902
906
Query Entry
NASA Thesaurus
(For thesaurus terms, cut/paste & separate by period)
apparatus for connecting a gas pressure source to several beer kegs in series
Display Options
Search
Sort Results by:
Claims
Relevance
Add results to my canvas
904
903
Analyze Query
Copyright © 1997 Manning & Napier Information Services
All Rights Reserved. MAPIT is a trademark of Manning & Napier Information Services.
Document Done

FIG. 9C

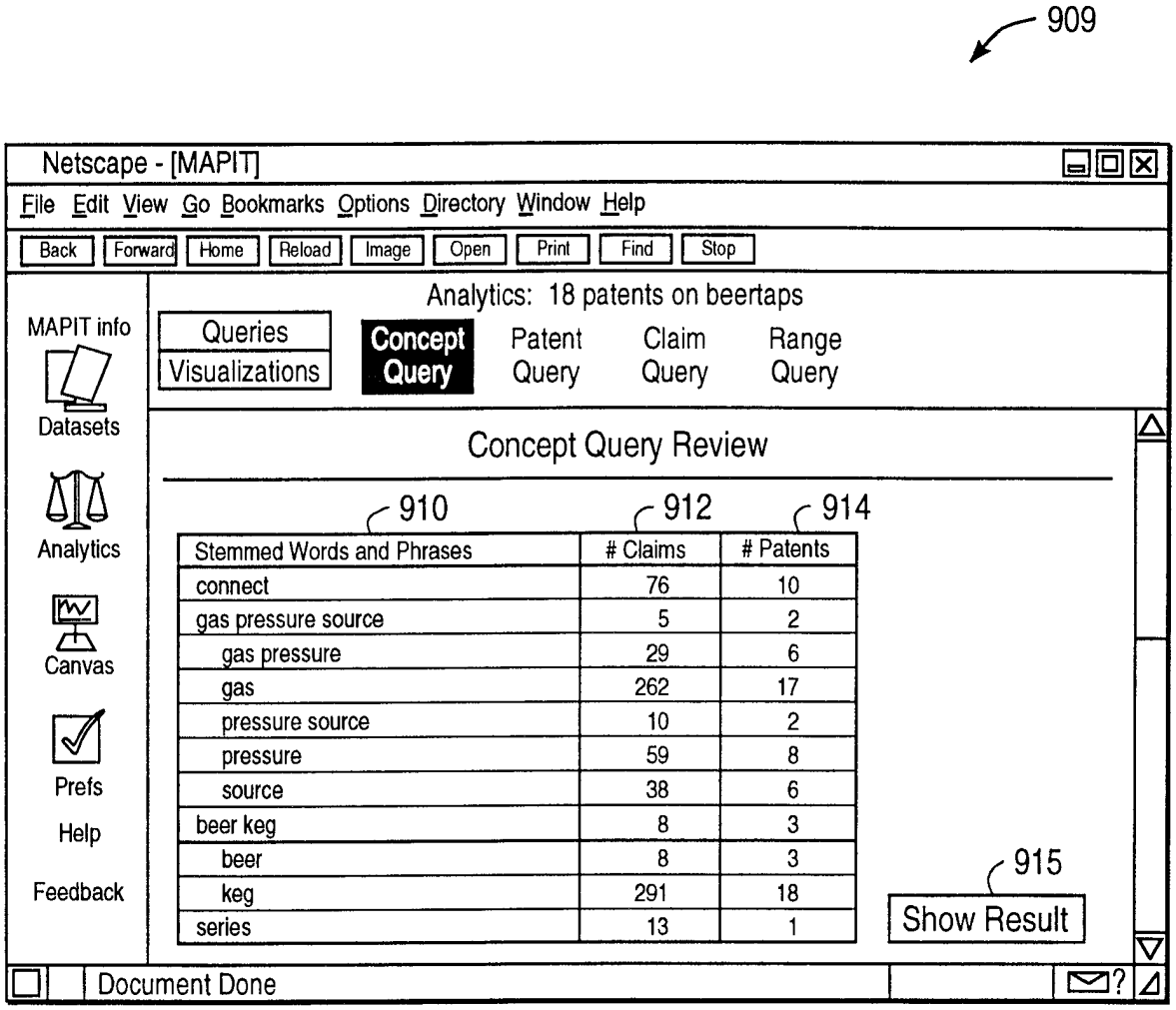

909
Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
Concept Query Review
910
912
914
Stemmed Words and Phrases
Claims
Patents
connect 76 10
gas pressure source 5 2
gas pressure 29 6
gas 262 17
pressure source 10 2
pressure 59 8
source 38 6
beer keg 8 3
beer 8 3
keg 291 18
series 13 1
915
Show Result
Document Done

FIG. 9D

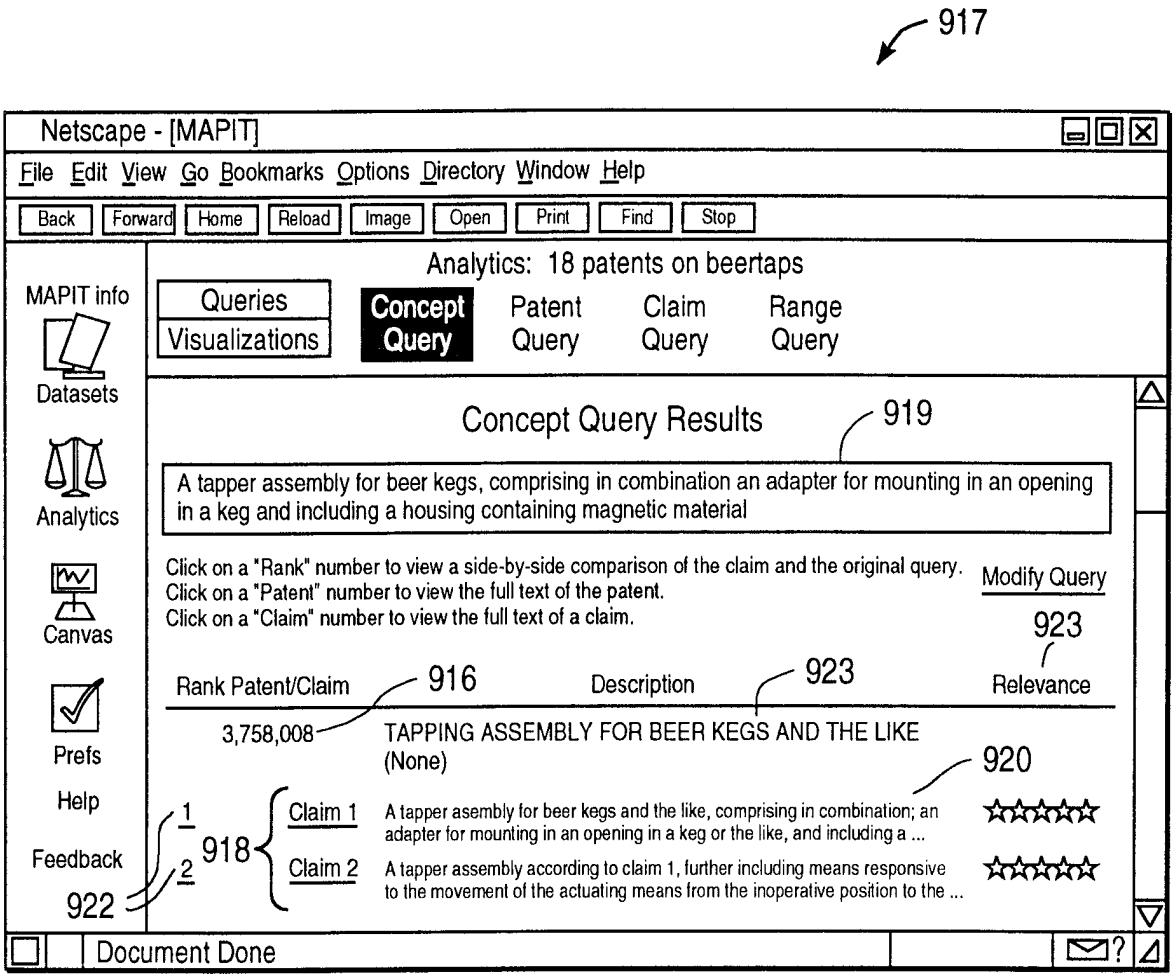

917
Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
922
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
Concept Query Results
919
A tapper assembly for beer kegs, comprising in combination an adapter for mounting in an opening in a keg and including a housing containing magnetic material
Click on a "Rank" number to view a side-by-side comparison of the claim and the original query.
Click on a "Patent" number to view the full text of the patent.
Click on a "Claim" number to view the full text of a claim.
Modify Query
923
Rank Patent/Claim
916
Description
923
Relevance
3,758,008
TAPPING ASSEMBLY FOR BEER KEGS AND THE LIKE
(None)
920
1
2
918
Claim 1
A tapper asembly for beer kegs and the like, comprising in combination; an adapter for mounting in an opening in a keg or the like, and including a ...
Claim 2
A tapper assembly according to claim 1, further including means responsive to the movement of the actuating means from the inoperative position to the ...
Document Done

FIG. 9E

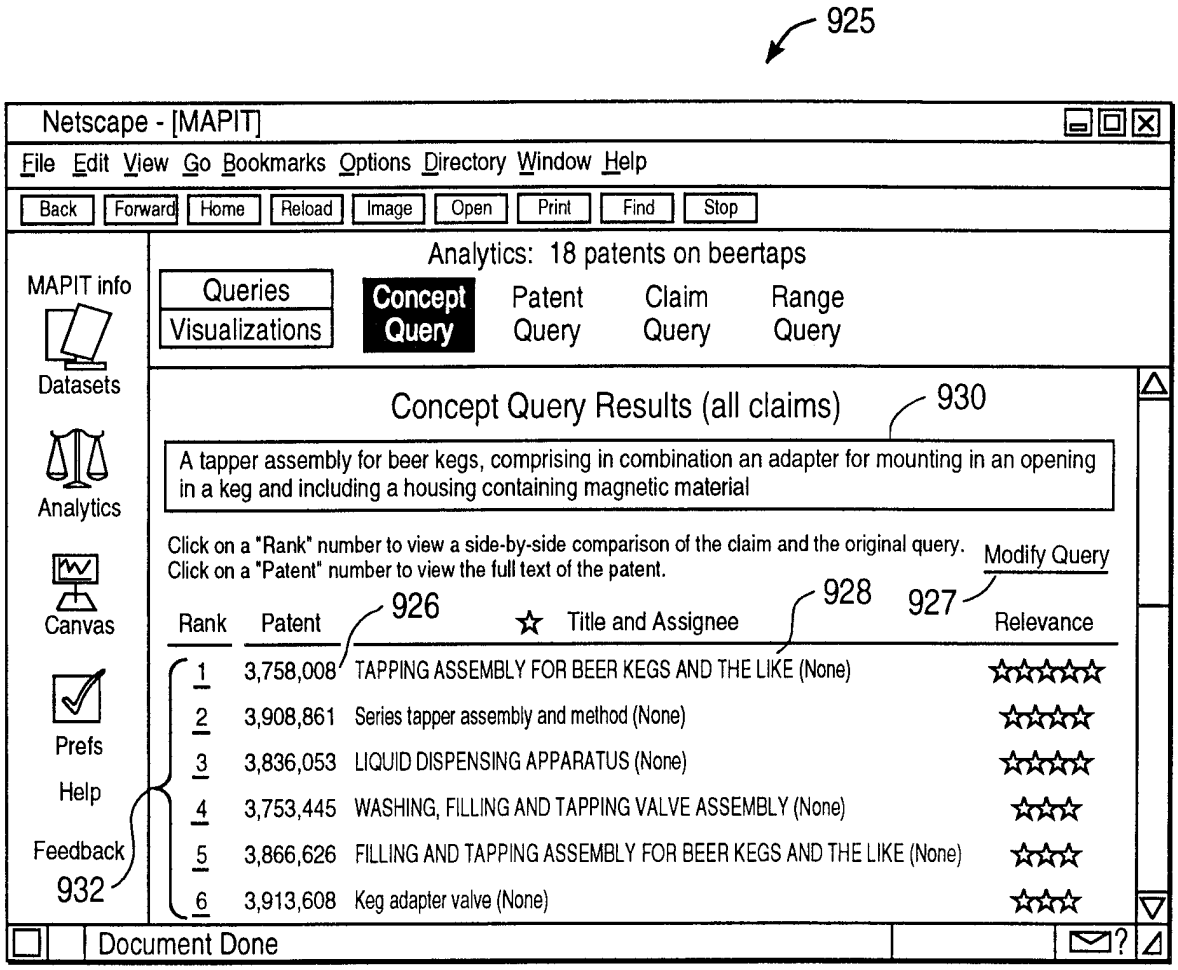
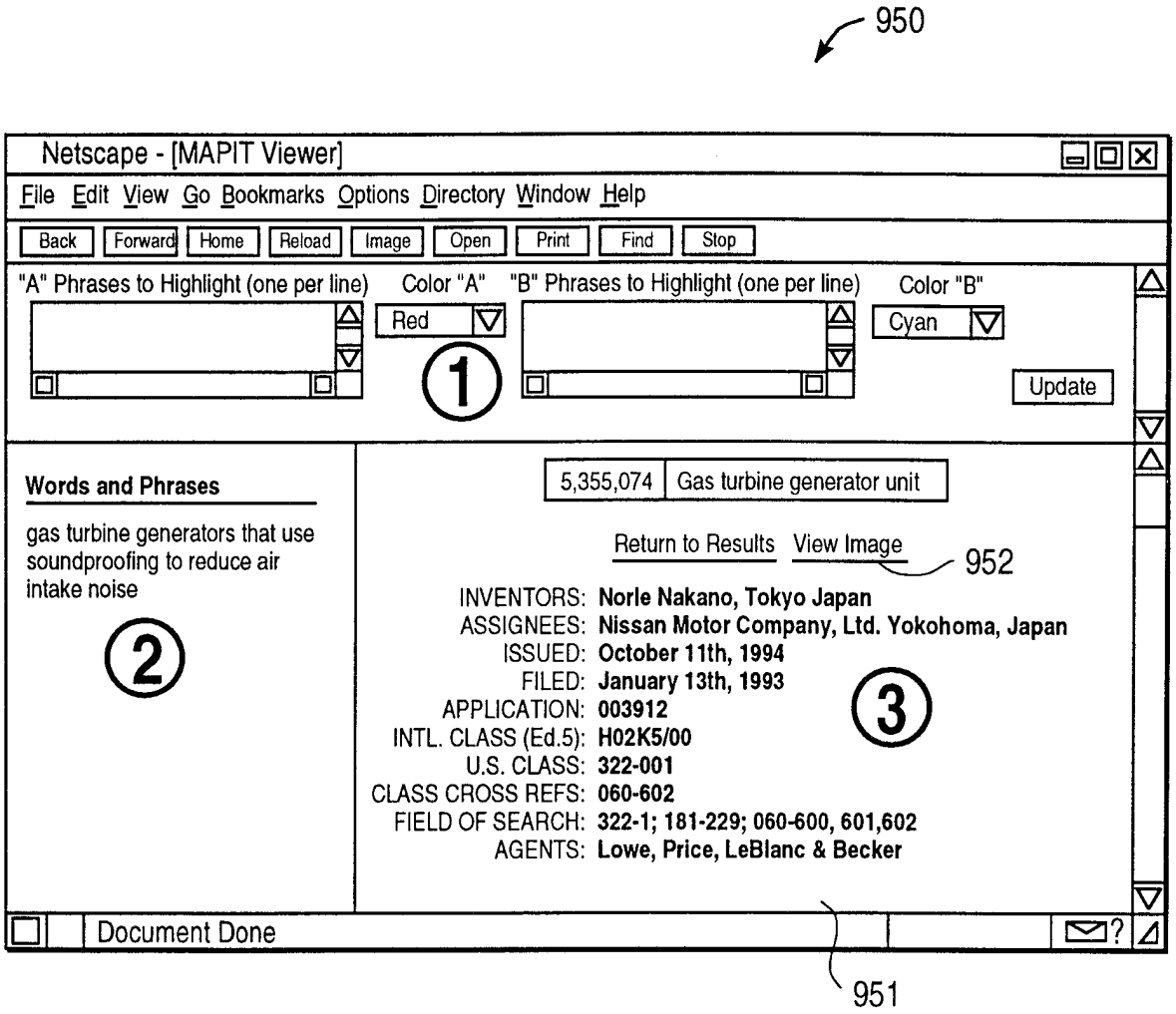
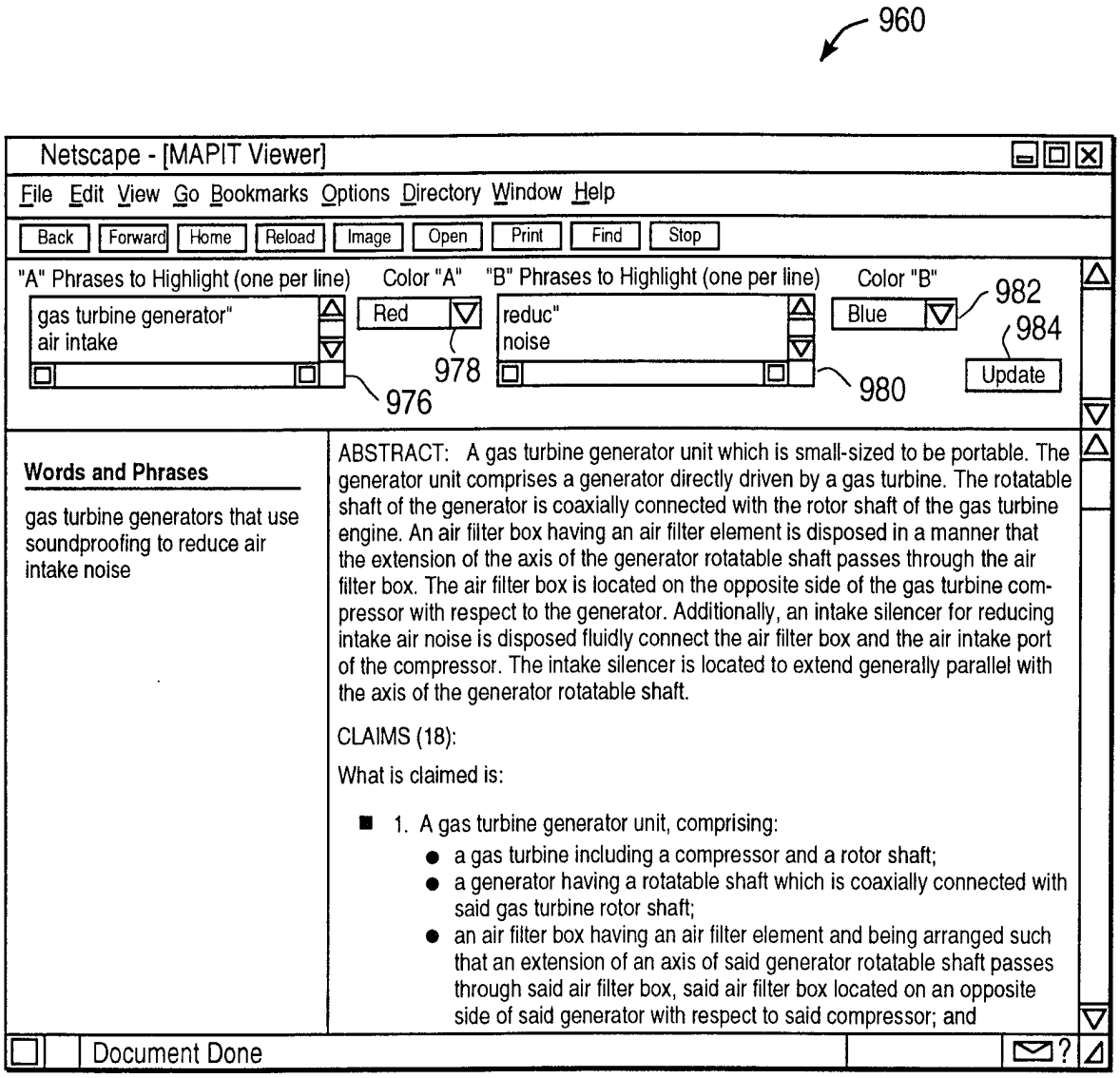

925
Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
932
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
Concept Query Results (all claims)
930
A tapper assembly for beer kegs, comprising in combination an adapter for mounting in an opening in a keg and including a housing containing magnetic material
Click on a "Rank" number to view a side-by-side comparison of the claim and the original query.
Click on a "Patent" number to view the full text of the patent.
Modify Query
926
928
927
Rank Patent ☆ Title and Assignee Relevance
1 3,758,008 TAPPING ASSEMBLY FOR BEER KEGS AND THE LIKE (None)
2 3,908,861 Series tapper assembly and method (None)
3 3,836,053 LIQUID DISPENSING APPARATUS (None)
4 3,753,445 WASHING, FILLING AND TAPPING VALVE ASSEMBLY (None)
5 3,866,626 FILLING AND TAPPING ASSEMBLY FOR BEER KEGS AND THE LIKE (None)
6 3,913,608 Keg adapter valve (None)
Document Done

FIG. 9F



The MAPIT Main Window

FIG. 9G



Textual Analysis with the Viewer

Figure 3.2:
The MAPIT Viewer window showing text highlighted in two colors.

FIG. 9H

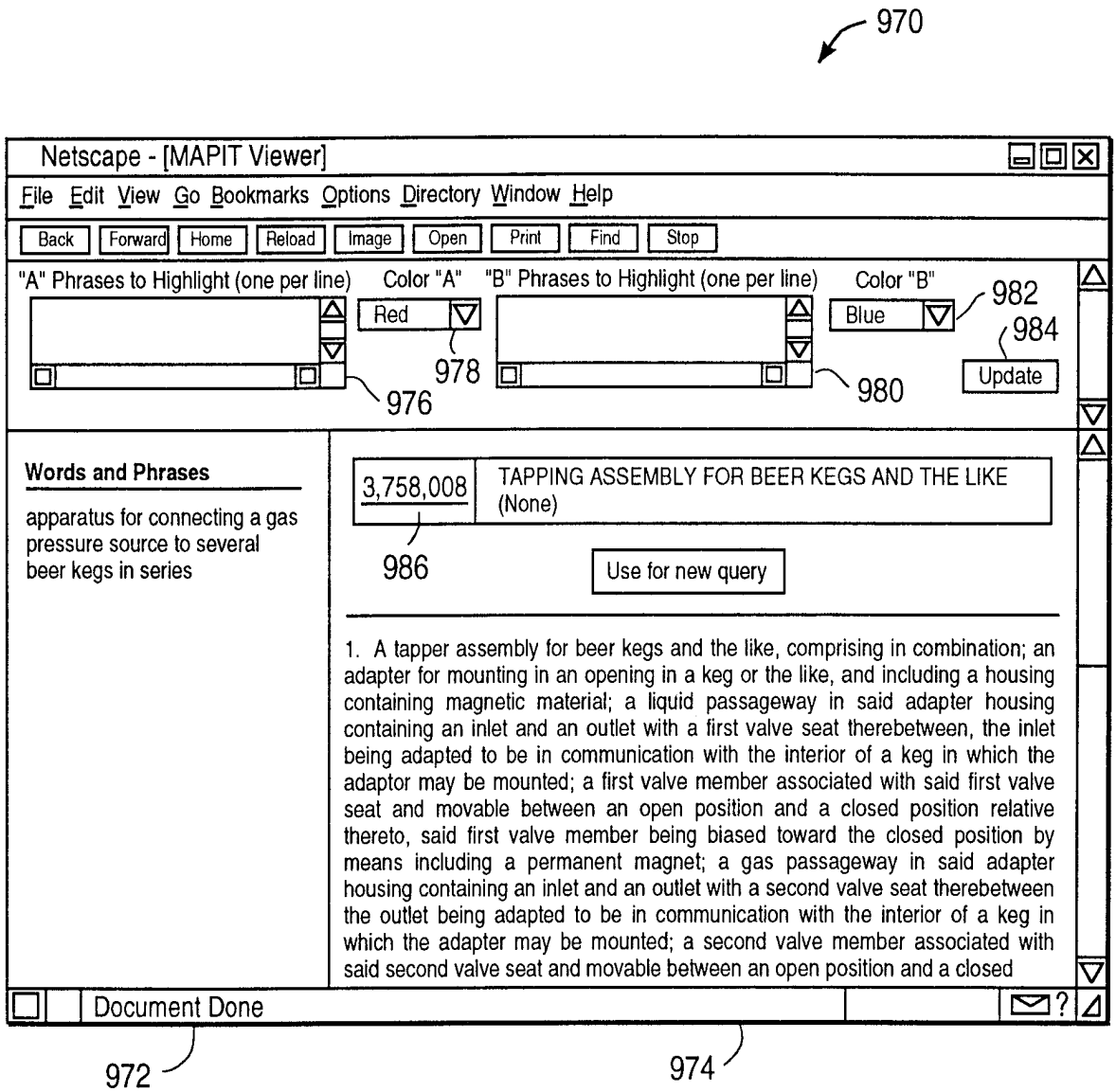

970
Netscape - [MAPIT Viewer]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
"A" Phrases to Highlight (one per line)
Color "A"
Red
"B" Phrases to Highlight (one per line)
Color "B"
Blue
982
984
Update
978
976
980
Words and Phrases
apparatus for connecting a gas pressure source to several beer kegs in series
3,758,008
TAPPING ASSEMBLY FOR BEER KEGS AND THE LIKE
(None)
986
Use for new query
1. A tapper assembly for beer kegs and the like, comprising in combination; an adapter for mounting in an opening in a keg or the like, and including a housing containing magnetic material; a liquid passageway in said adapter housing containing an inlet and an outlet with a first valve seat therebetween, the inlet being adapted to be in communication with the interior of a keg in which the adaptor may be mounted; a first valve member associated with said first valve seat and movable between an open position and a closed position relative thereto, said first valve member being biased toward the closed position by means including a permanent magnet; a gas passageway in said adapter housing containing an inlet and an outlet with a second valve seat therebetween the outlet being adapted to be in communication with the interior of a keg in which the adapter may be mounted; a second valve member associated with said second valve seat and movable between an open position and a closed
Document Done
972
974

FIG. 9I

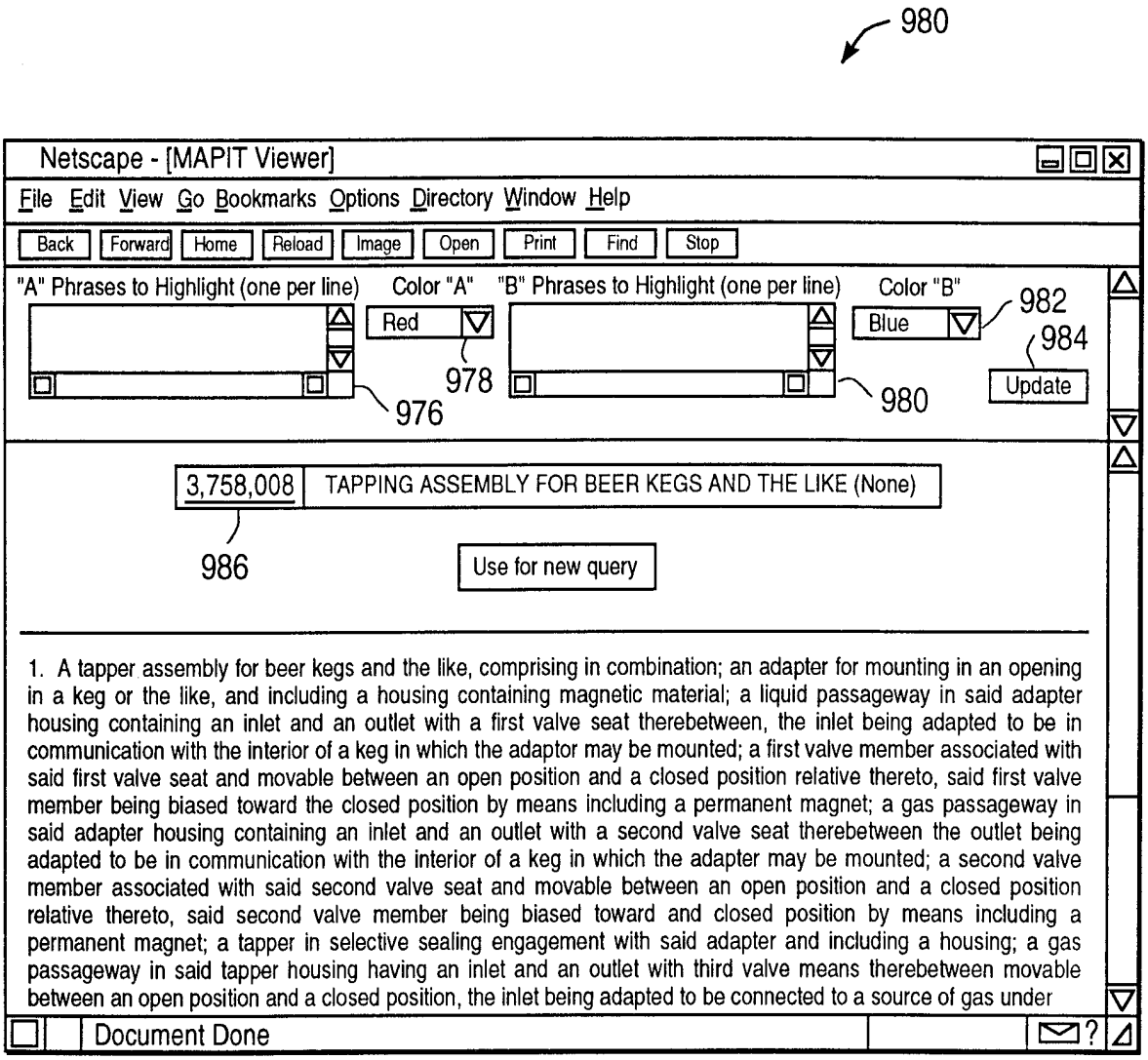

980
Netscape - [MAPIT Viewer]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
"A" Phrases to Highlight (one per line)
Color "A"
Red
"B" Phrases to Highlight (one per line)
Color "B"
Blue
982
984
Update
978
976
980
3,758,008
TAPPING ASSEMBLY FOR BEER KEGS AND THE LIKE (None)
986
Use for new query
1. A tapper assembly for beer kegs and the like, comprising in combination; an adapter for mounting in an opening in a keg or the like, and including a housing containing magnetic material; a liquid passageway in said adapter housing containing an inlet and an outlet with a first valve seat therebetween, the inlet being adapted to be in communication with the interior of a keg in which the adaptor may be mounted; a first valve member associated with said first valve seat and movable between an open position and a closed position relative thereto, said first valve member being biased toward the closed position by means including a permanent magnet; a gas passageway in said adapter housing containing an inlet and an outlet with a second valve seat therebetween the outlet being adapted to be in communication with the interior of a keg in which the adapter may be mounted; a second valve member associated with said second valve seat and movable between an open position and a closed position relative thereto, said second valve member being biased toward and closed position by means including a permanent magnet; a tapper in selective sealing engagement with said adapter and including a housing; a gas passageway in said tapper housing having an inlet and an outlet with third valve means therebetween movable between an open position and a closed position, the inlet being adapted to be connected to a source of gas under
Document Done

FIG. 9J

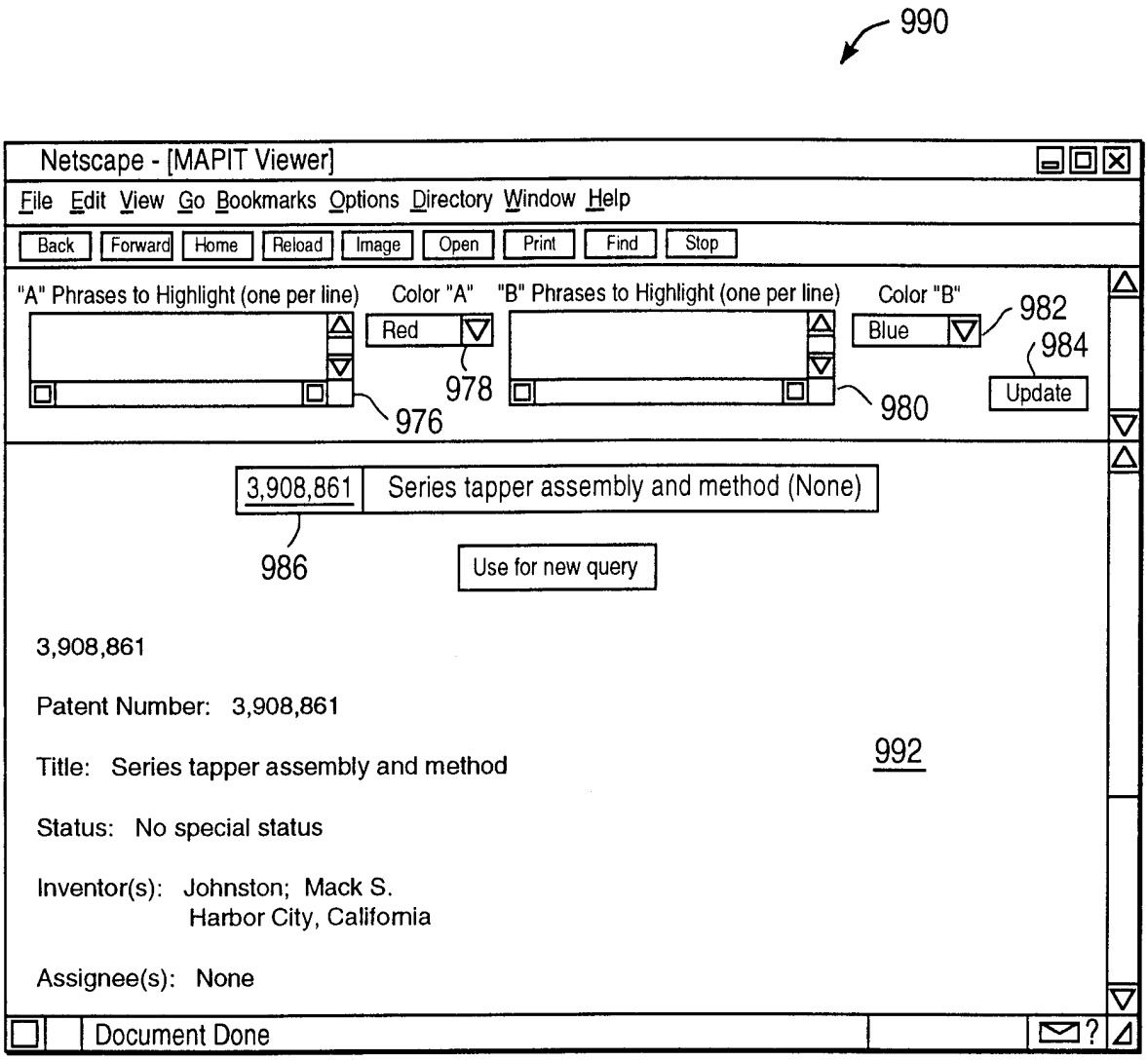

990
Netscape - [MAPIT Viewer]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
"A" Phrases to Highlight (one per line)
Color "A"
Red
"B" Phrases to Highlight (one per line)
Color "B"
Blue
982
984
Update
978
976
980
3,908,861
Series tapper assembly and method (None)
986
Use for new query
3,908,861
Patent Number: 3,908,861
Title: Series tapper assembly and method
992
Status: No special status
Inventor(s): Johnston; Mack S.
Harbor City, California
Assignee(s): None
Document Done

FIG. 9K

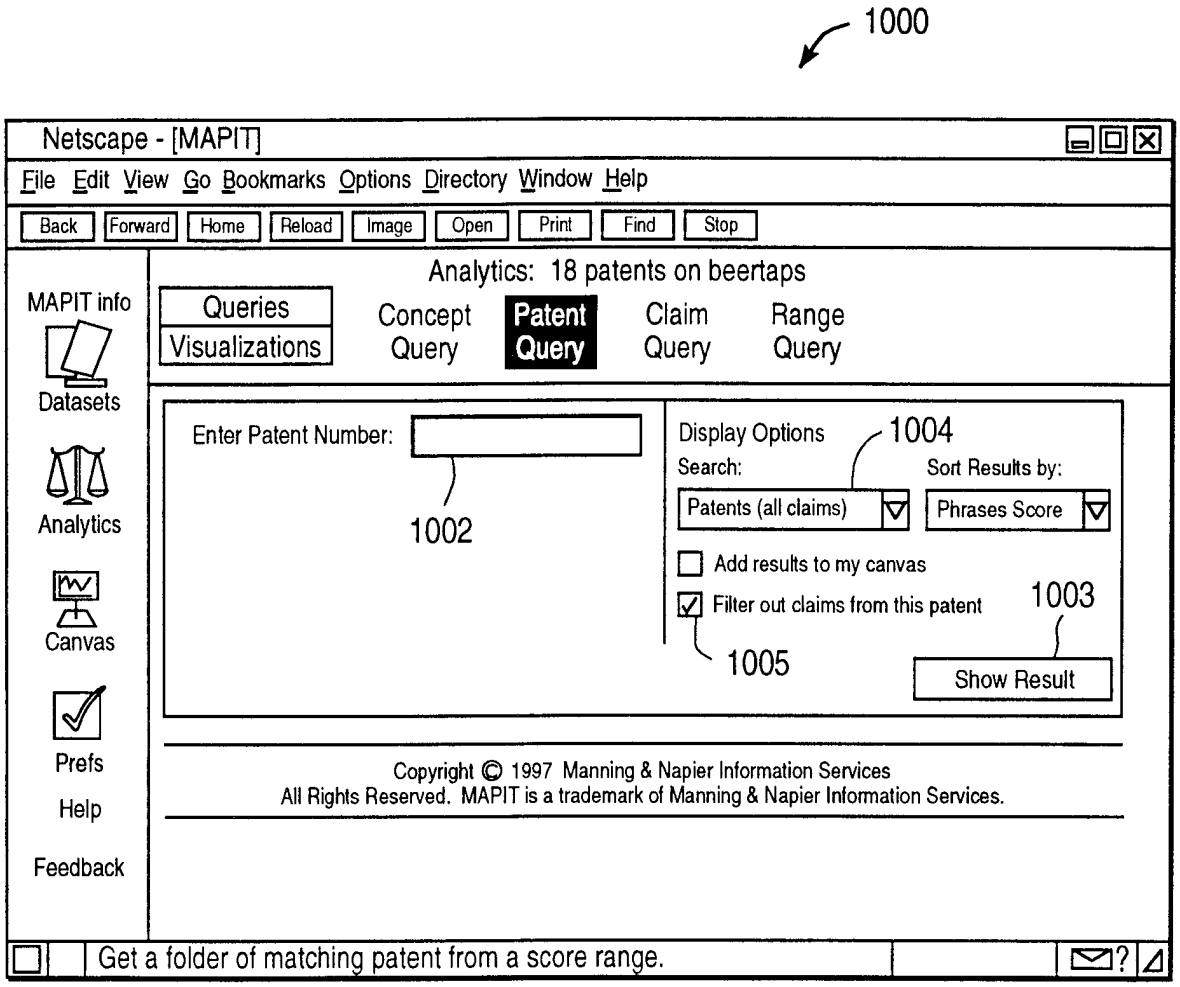

1000
Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
Enter Patent Number:
1002
Display Options
1004
Search:
Sort Results by:
Patents (all claims)
Phrases Score
Add results to my canvas
Filter out claims from this patent
1003
1005
Show Result
Copyright © 1997 Manning & Napier Information Services
All Rights Reserved. MAPIT is a trademark of Manning & Napier Information Services.
Get a folder of matching patent from a score range.

FIG. 10A

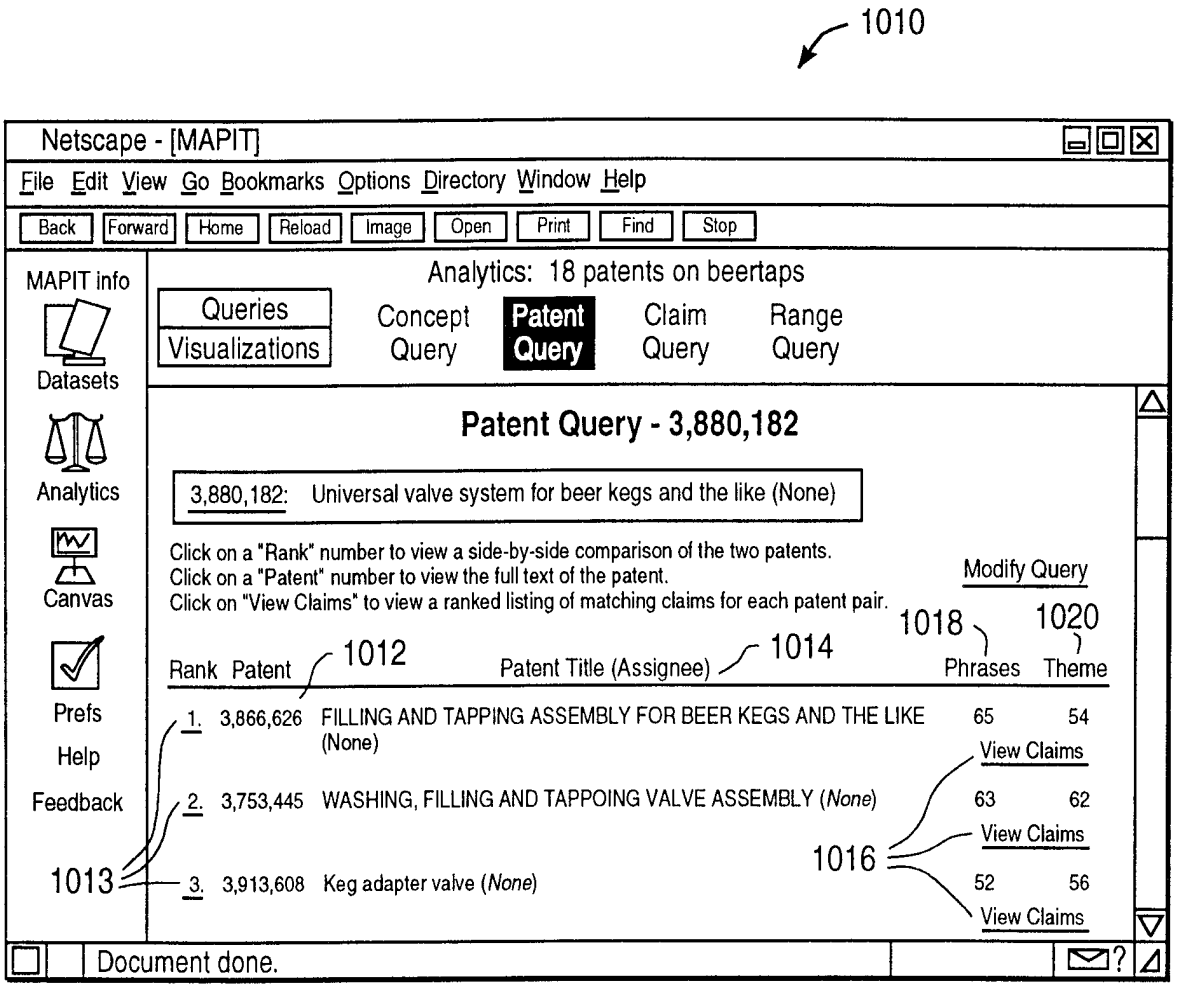

1010
Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
Patent Query - 3,880,182
3,880,182: Universal valve system for beer kegs and the like (None)
Click on a "Rank" number to view a side-by-side comparison of the two patents.
Click on a "Patent" number to view the full text of the patent.
Click on "View Claims" to view a ranked listing of matching claims for each patent pair.
Modify Query
1018
1020
Rank Patent
1012
Patent Title (Assignee)
1014
Phrases
Theme
1. 3,866,626 FILLING AND TAPPING ASSEMBLY FOR BEER KEGS AND THE LIKE (None) 65 54 View Claims
2. 3,753,445 WASHING, FILLING AND TAPPOING VALVE ASSEMBLY (None) 63 62 View Claims
3. 3,913,608 Keg adapter valve (None) 52 56 View Claims
1013
1016
Document done.

FIG. 10B

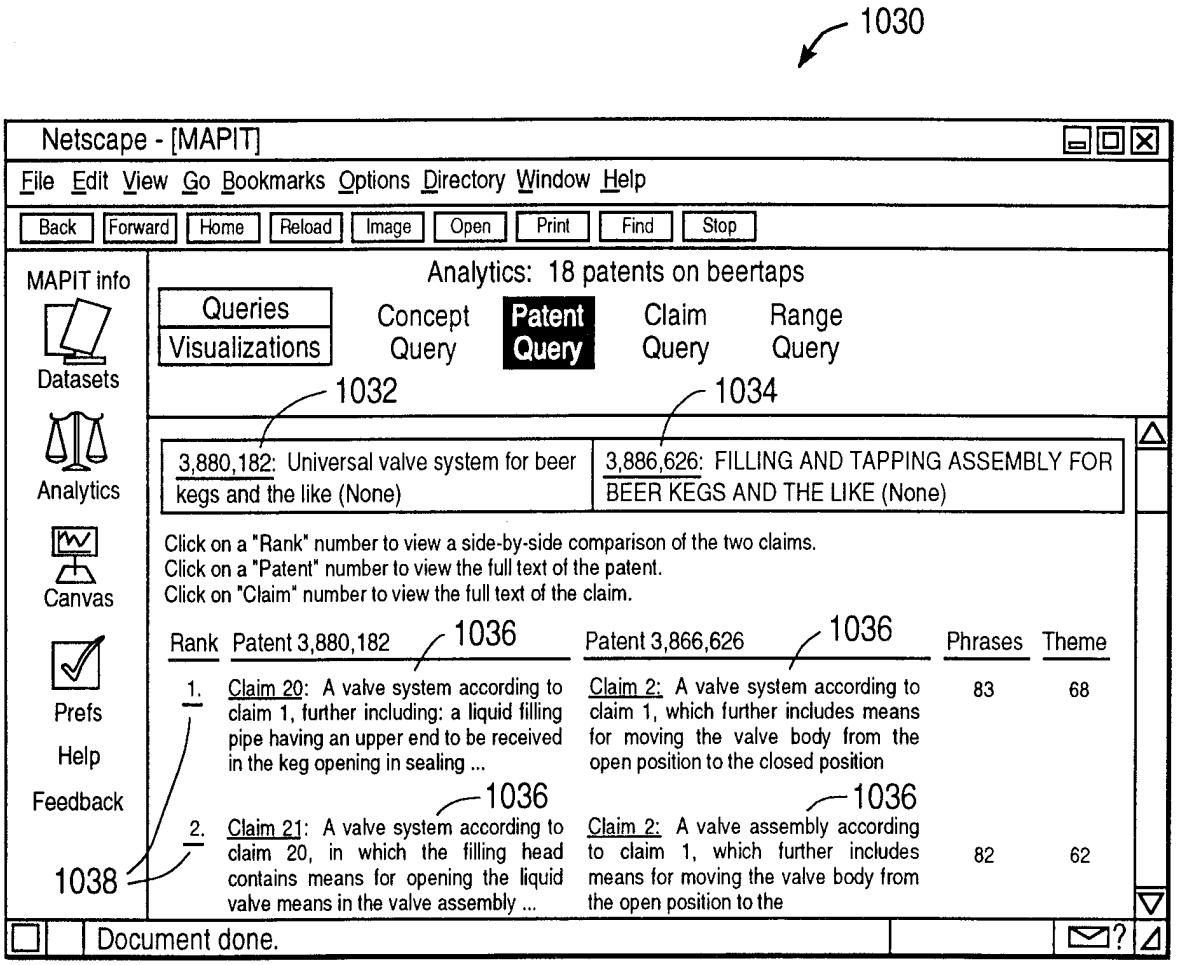

1030
Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
1038
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
1032
1034
3,880,182: Universal valve system for beer kegs and the like (None)
3,886,626: FILLING AND TAPPING ASSEMBLY FOR BEER KEGS AND THE LIKE (None)
Click on a "Rank" number to view a side-by-side comparison of the two claims.
Click on a "Patent" number to view the full text of the patent.
Click on "Claim" number to view the full text of the claim.
Rank
Patent 3,880,182
1036
Patent 3,866,626
1036
Phrases
Theme
1. Claim 20: A valve system according to claim 1, further including: a liquid filling pipe having an upper end to be received in the keg opening in sealing ...
Claim 2: A valve system according to claim 1, which further includes means for moving the valve body from the open position to the closed position
83
68
2. Claim 21: A valve system according to claim 20, in which the filling head contains means for opening the liquid valve means in the valve assembly ...
1036
Claim 2: A valve assembly according to claim 1, which further includes means for moving the valve body from the open position to the
1036
82
62
Document done.

FIG. 10C

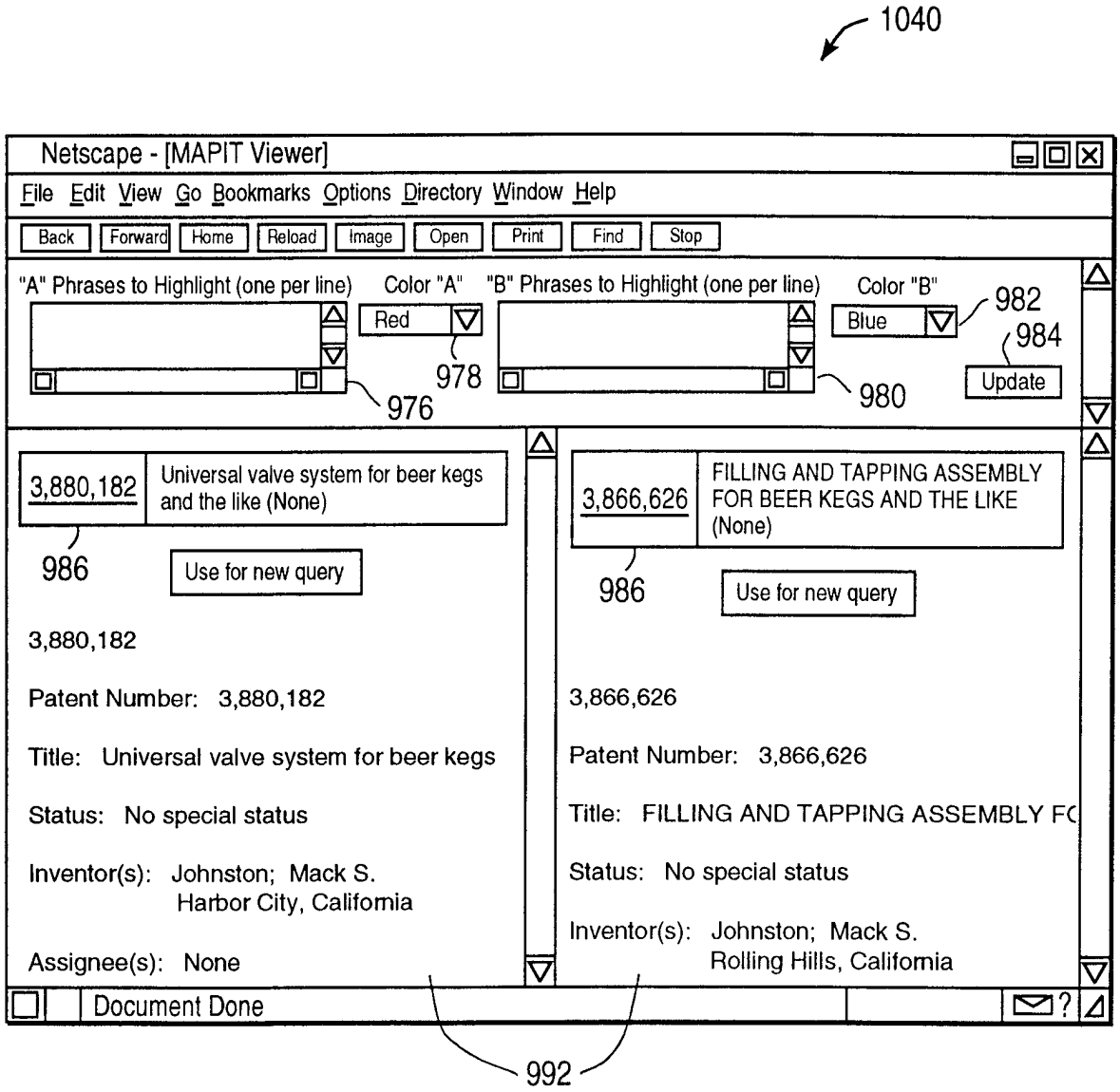

1040
Netscape - [MAPIT Viewer]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
"A" Phrases to Highlight (one per line)
Color "A"
Red
978
976
"B" Phrases to Highlight (one per line)
Color "B"
Blue
982
984
980
Update
3,880,182
Universal valve system for beer kegs and the like (None)
986
Use for new query
3,880,182
Patent Number: 3,880,182
Title: Universal valve system for beer kegs
Status: No special status
Inventor(s): Johnston; Mack S.
Harbor City, California
Assignee(s): None
3,866,626
FILLING AND TAPPING ASSEMBLY FOR BEER KEGS AND THE LIKE (None)
986
Use for new query
3,866,626
Patent Number: 3,866,626
Title: FILLING AND TAPPING ASSEMBLY FC
Status: No special status
Inventor(s): Johnston; Mack S.
Rolling Hills, California
Document Done
992

FIG. 10D

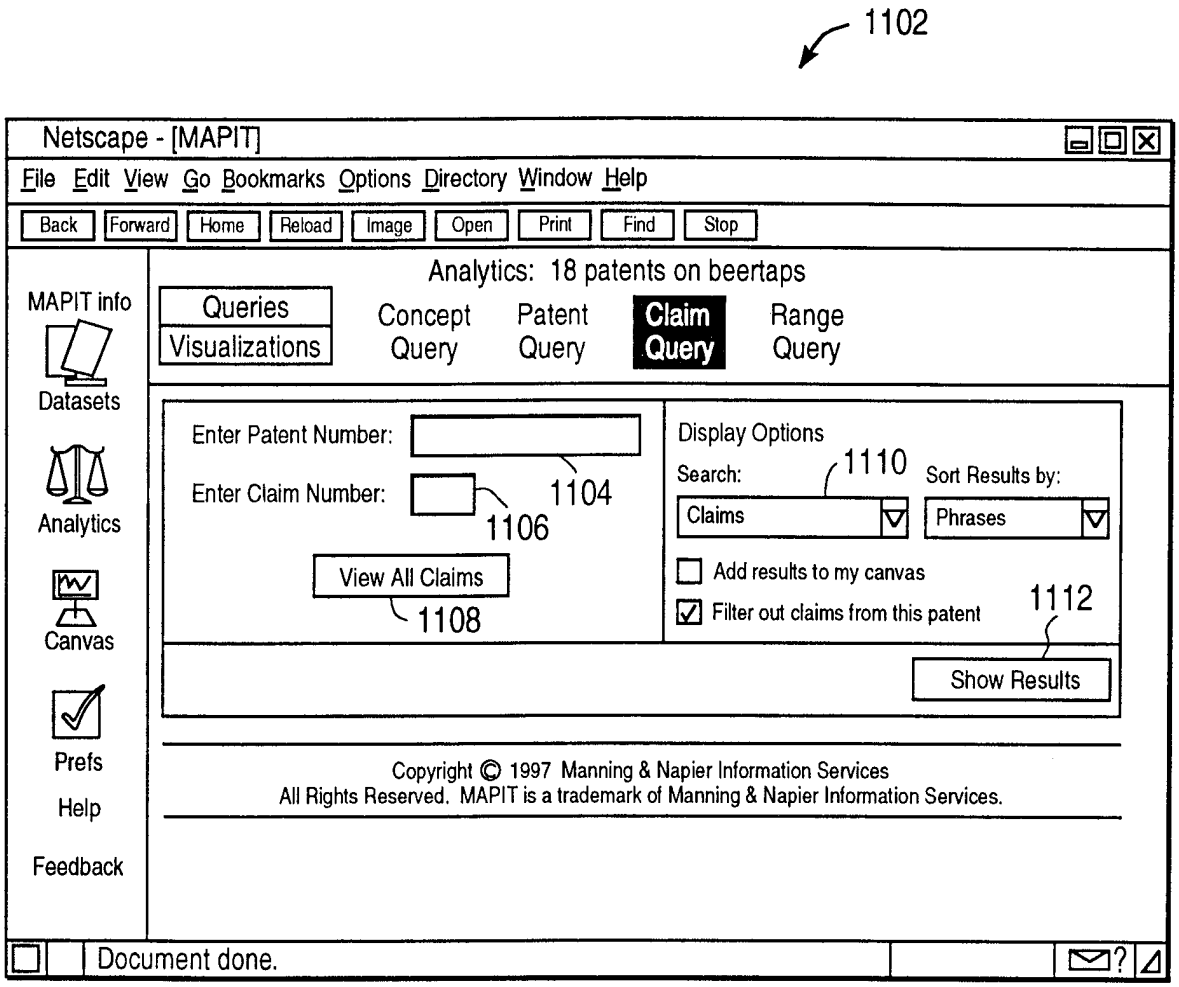

1102
Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
Enter Patent Number:
Enter Claim Number:
1104
1106
View All Claims
1108
Display Options
Search:
1110
Sort Results by:
Claims
Phrases
Add results to my canvas
Filter out claims from this patent
1112
Show Results
Copyright © 1997 Manning & Napier Information Services
All Rights Reserved. MAPIT is a trademark of Manning & Napier Information Services.
Document done.

FIG. 11A

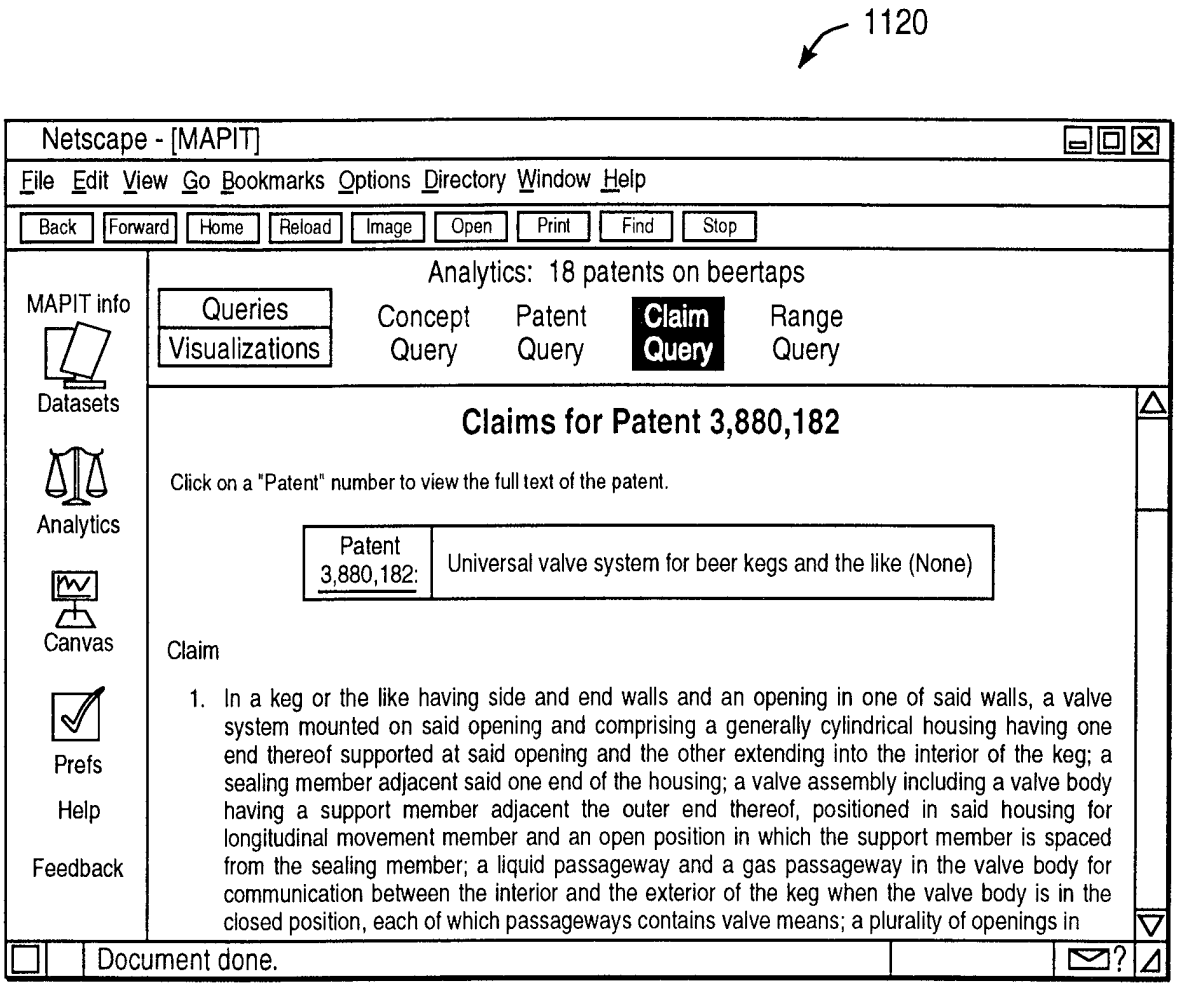

1120
Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
Claims for Patent 3,880,182
Click on a "Patent" number to view the full text of the patent.
Patent 3,880,182:
Universal valve system for beer kegs and the like (None)
Claim
1. In a keg or the like having side and end walls and an opening in one of said walls, a valve system mounted on said opening and comprising a generally cylindrical housing having one end thereof supported at said opening and the other extending into the interior of the keg; a sealing member adjacent said one end of the housing; a valve assembly including a valve body having a support member adjacent the outer end thereof, positioned in said housing for longitudinal movement member and an open position in which the support member is spaced from the sealing member; a liquid passageway and a gas passageway in the valve body for communication between the interior and the exterior of the keg when the valve body is in the closed position, each of which passageways contains valve means; a plurality of openings in
Document done.

FIG. 11B

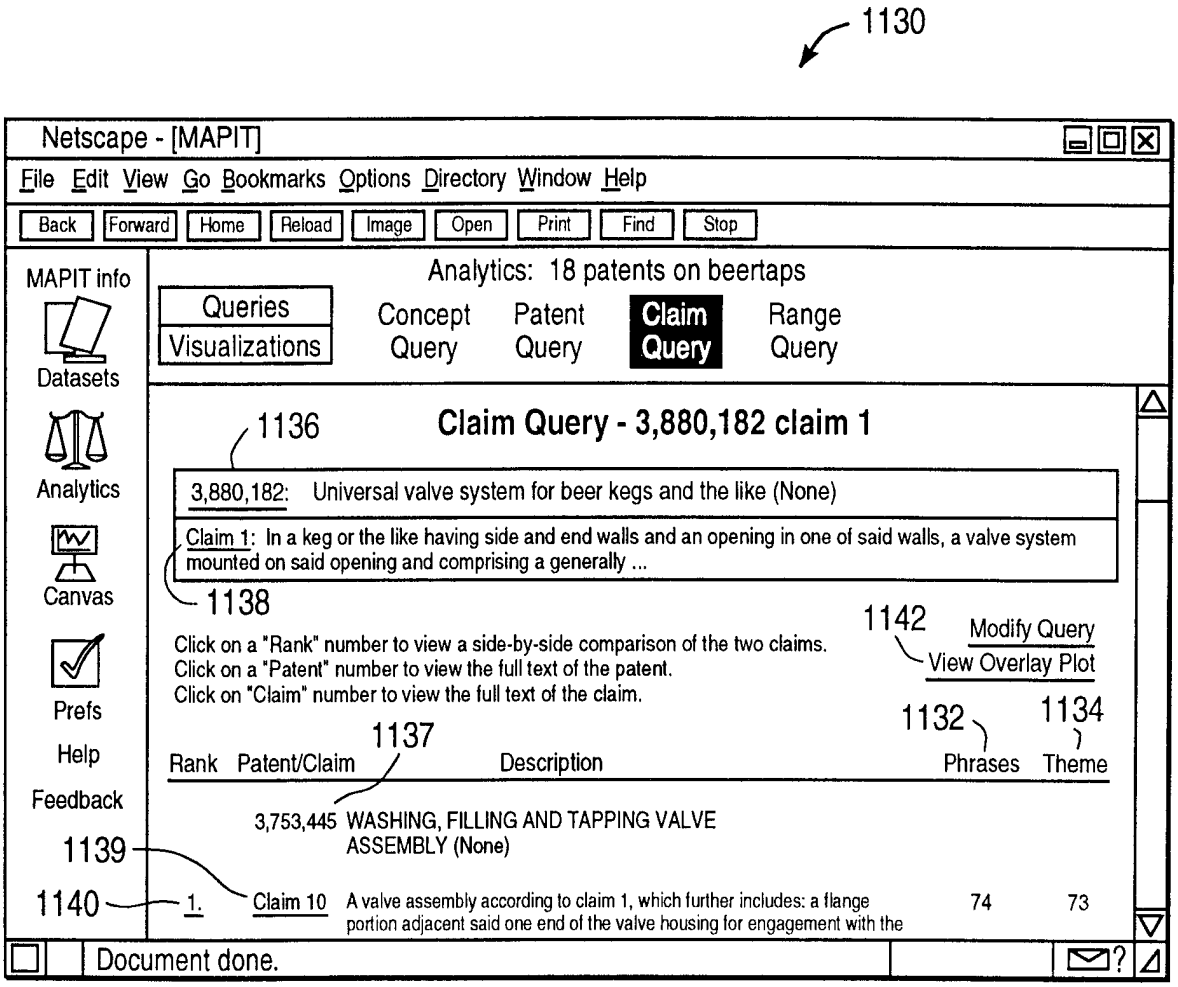

1130
Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
1136
Claim Query - 3,880,182 claim 1
3,880,182: Universal valve system for beer kegs and the like (None)
Claim 1: In a keg or the like having side and end walls and an opening in one of said walls, a valve system mounted on said opening and comprising a generally ...
1138
1142
Modify Query
View Overlay Plot
Click on a "Rank" number to view a side-by-side comparison of the two claims.
Click on a "Patent" number to view the full text of the patent.
Click on "Claim" number to view the full text of the claim.
1132
1134
1137
Rank Patent/Claim Description Phrases Theme
3,753,445 WASHING, FILLING AND TAPPING VALVE ASSEMBLY (None)
1139
1140
1. Claim 10 A valve assembly according to claim 1, which further includes: a flange portion adjacent said one end of the valve housing for engagement with the 74 73
Document done.

FIG. 11C

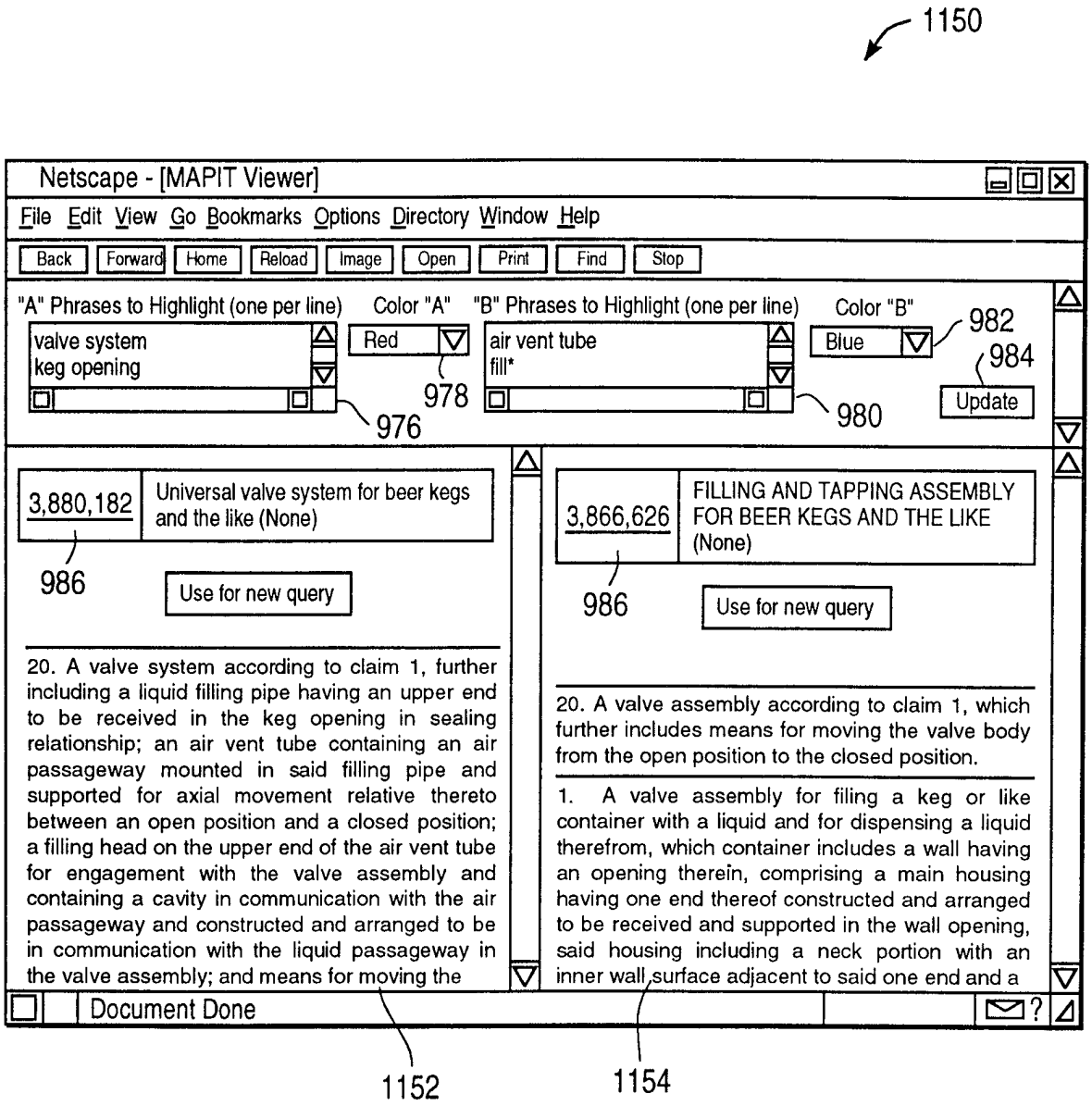

1150
Netscape - [MAPIT Viewer]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
"A" Phrases to Highlight (one per line)
valve system
keg opening
Color "A"
Red
978
976
"B" Phrases to Highlight (one per line)
air vent tube
fill*
980
Color "B"
Blue
982
984
Update
3,880,182
Universal valve system for beer kegs and the like (None)
986
Use for new query
20. A valve system according to claim 1, further including a liquid filling pipe having an upper end to be received in the keg opening in sealing relationship; an air vent tube containing an air passageway mounted in said filling pipe and supported for axial movement relative thereto between an open position and a closed position; a filling head on the upper end of the air vent tube for engagement with the valve assembly and containing a cavity in communication with the air passageway and constructed and arranged to be in communication with the liquid passageway in the valve assembly; and means for moving the
3,866,626
FILLING AND TAPPING ASSEMBLY FOR BEER KEGS AND THE LIKE (None)
986
Use for new query
20. A valve assembly according to claim 1, which further includes means for moving the valve body from the open position to the closed position.
1. A valve assembly for filing a keg or like container with a liquid and for dispensing a liquid therefrom, which container includes a wall having an opening therein, comprising a main housing having one end thereof constructed and arranged to be received and supported in the wall opening, said housing including a neck portion with an inner wall surface adjacent to said one end and a
Document Done
1152
1154

FIG. 11D

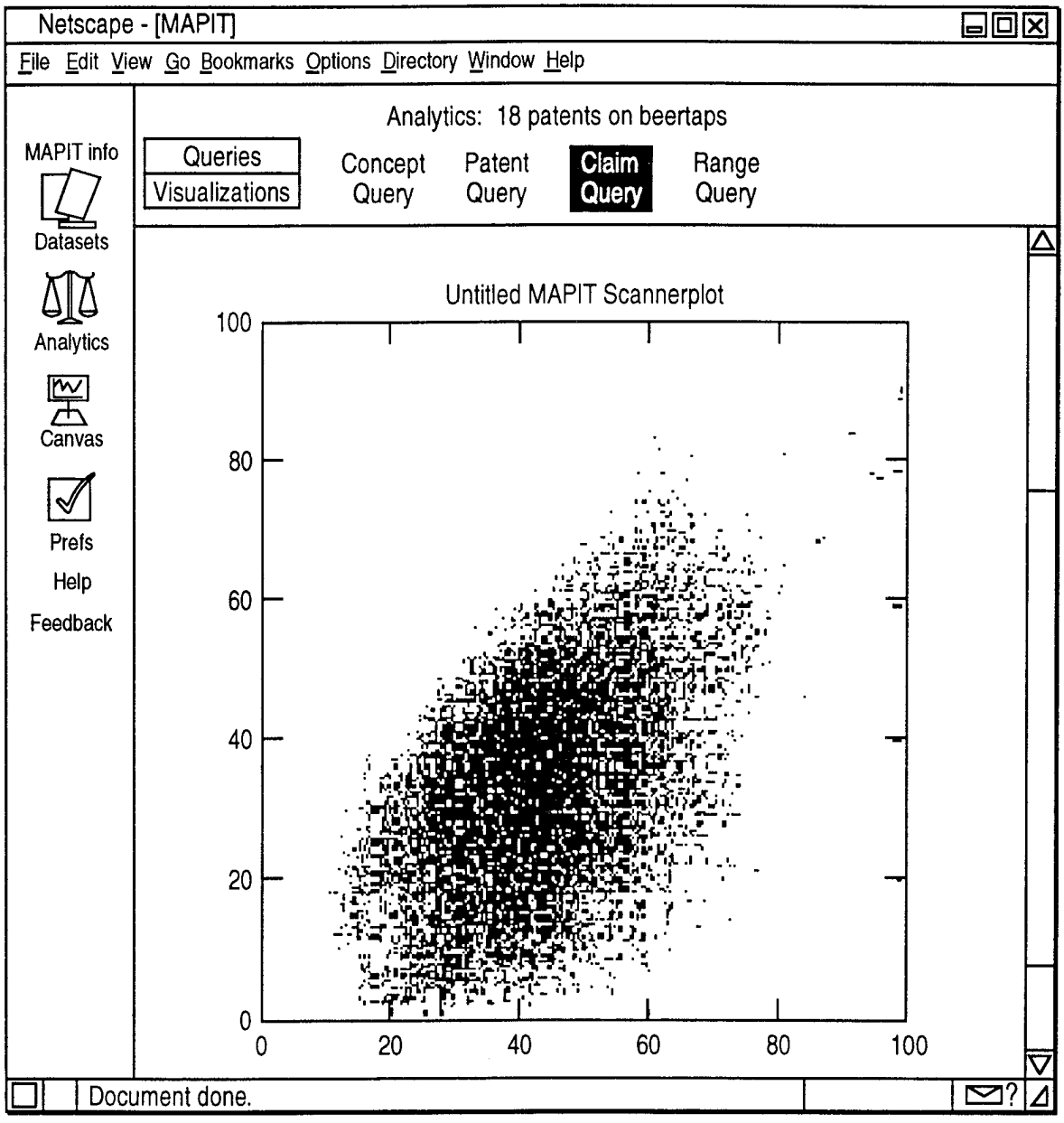

Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
Untitled MAPIT Scannerplot
100
80
60
40
20
0
0
20
40
60
80
100
Document done.

FIG. 11E

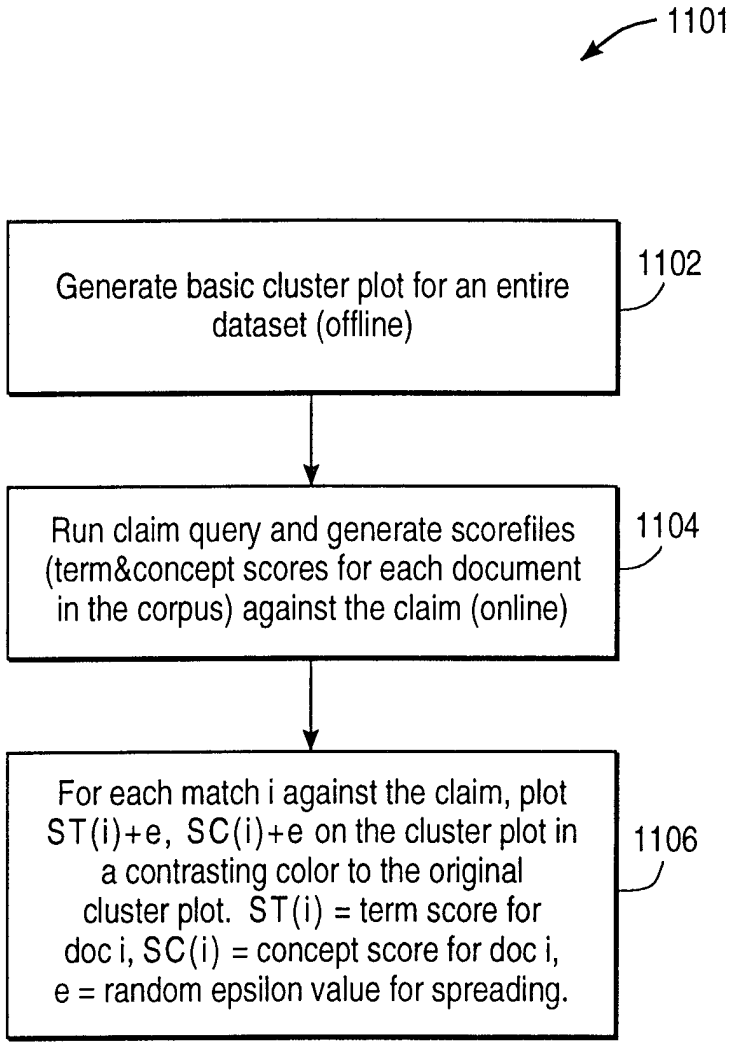

1101
Generate basic cluster plot for an entire dataset (offline)
1102
Run claim query and generate scorefiles (term&concept scores for each document in the corpus) against the claim (online)
1104
For each match i against the claim, plot ST(i)+e, SC(i)+e on the cluster plot in a contrasting color to the original cluster plot. ST(i) = term score for doc i, SC(i) = concept score for doc i, e = random epsilon value for spreading.
1106

FIG. 11F

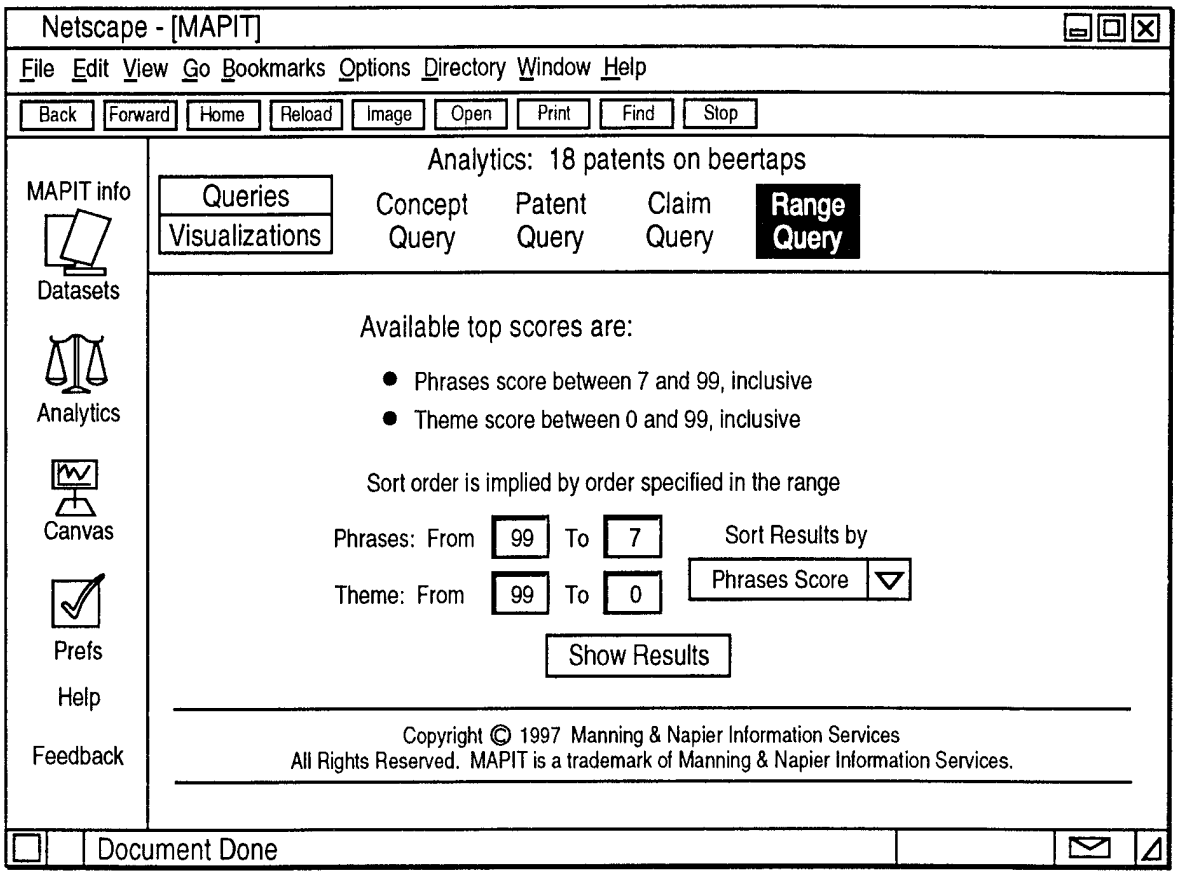

Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
Available top scores are:
Phrases score between 7 and 99, inclusive
Theme score between 0 and 99, inclusive
Sort order is implied by order specified in the range
Phrases: From 99 To 7
Theme: From 99 To 0
Sort Results by
Phrases Score
Show Results
Copyright © 1997 Manning & Napier Information Services
All Rights Reserved. MAPIT is a trademark of Manning & Napier Information Services.
Document Done

FIG. 12A

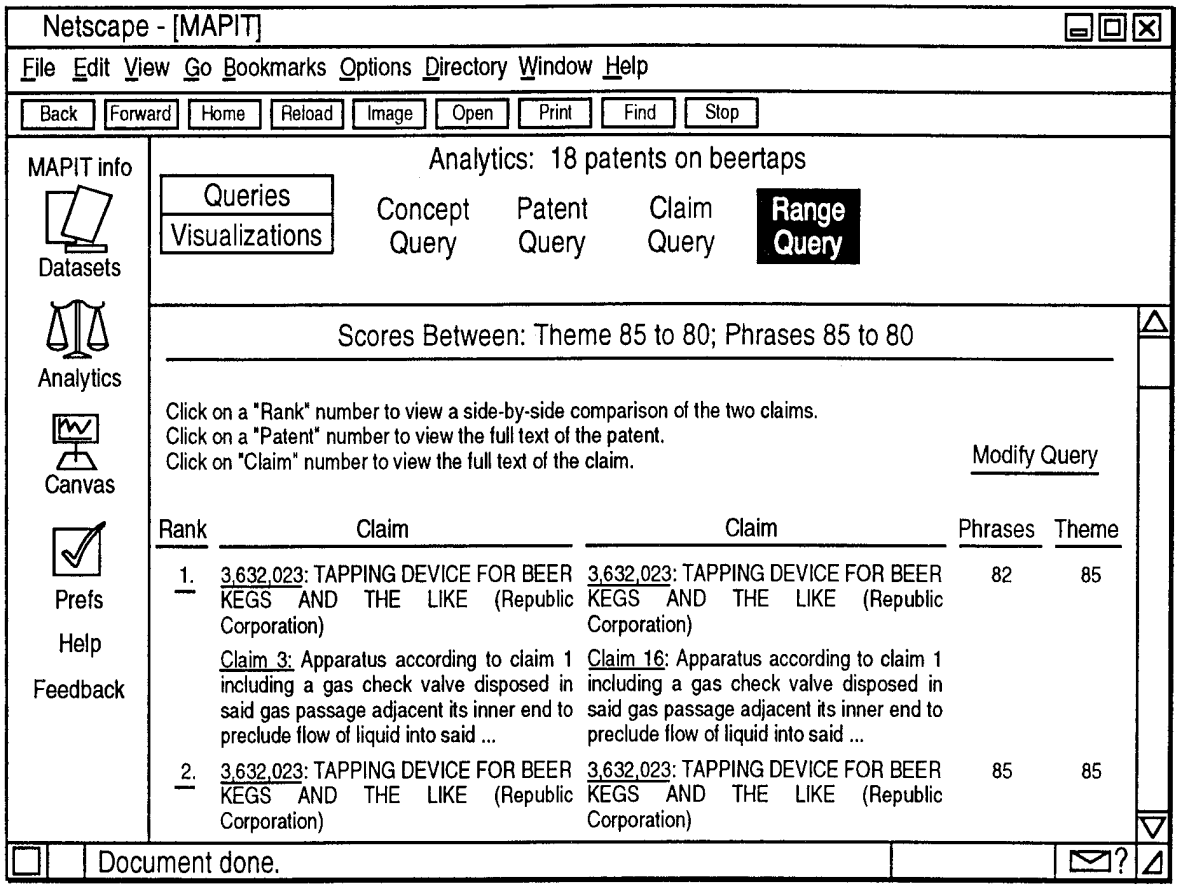

Netscape - [MAPIT]
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Image Open Print Find Stop
MAPIT info
Datasets
Analytics
Canvas
Prefs
Help
Feedback
Analytics: 18 patents on beertaps
Queries
Visualizations
Concept Query
Patent Query
Claim Query
Range Query
Scores Between: Theme 85 to 80; Phrases 85 to 80
Click on a "Rank" number to view a side-by-side comparison of the two claims.
Click on a "Patent" number to view the full text of the patent.
Click on "Claim" number to view the full text of the claim.
Modify Query
Rank
Claim
Claim
Phrases
Theme
1.
3,632,023: TAPPING DEVICE FOR BEER KEGS AND THE LIKE (Republic Corporation)
Claim 3: Apparatus according to claim 1 including a gas check valve disposed in said gas passage adjacent its inner end to preclude flow of liquid into said ...
3,632,023: TAPPING DEVICE FOR BEER KEGS AND THE LIKE (Republic Corporation)
Claim 16: Apparatus according to claim 1 including a gas check valve disposed in said gas passage adjacent its inner end to preclude flow of liquid into said ...
82
85
2.
3,632,023: TAPPING DEVICE FOR BEER KEGS AND THE LIKE (Republic Corporation)
3,632,023: TAPPING DEVICE FOR BEER KEGS AND THE LIKE (Republic Corporation)
85
85
Document done.

FIG. 12B

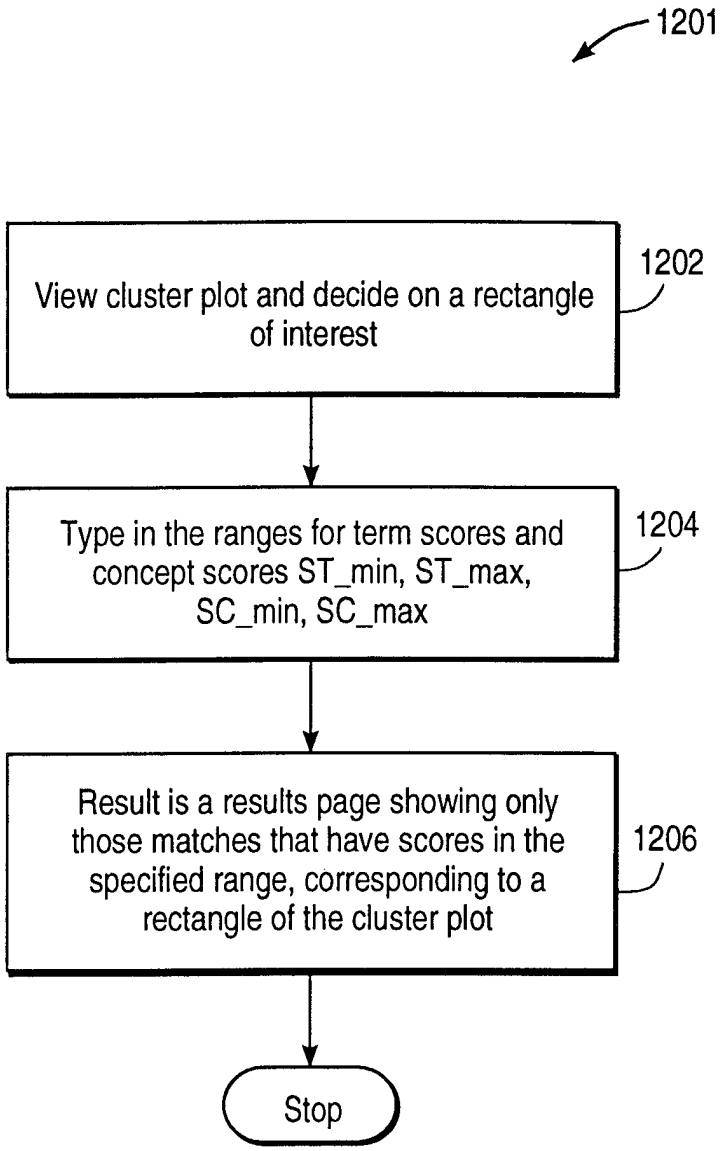

1201
View cluster plot and decide on a rectangle of interest
1202
Type in the ranges for term scores and concept scores ST_min, ST_max, SC_min, SC_max
1204
Result is a results page showing only those matches that have scores in the specified range, corresponding to a rectangle of the cluster plot
1206
Stop

FIG. 12C

MAPIT: Select Base Claim

To do claim analysis with MAPIT, you must first select a base claim. MAPIT will then compare all other claims in the patent set against that bae claim.

You may enter a list of keywords to search for claims, or you may select a claim from a specific application

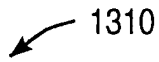
Select using words/phrases		Go (keywords)
Select from **all** applications	Go (all apps)	
Select from **specific** application	35196 4074324 4531161 4574319 4642678 4660092 4663661 4675755 4714963 4746988	Go (specified app)
Show only **scores** above	0.80 ○ 0.90 ○ 0.95 ◉	Go (high scores)

If you want to preview the full data set, click on one of the boxes below. "Merged" includes the text of referred claims; "Single" does not. "Compare" plots two different similarity measures on the same graph.

MERGED	2D	3D	Compare
SINGLE	2D	3D	Compare

PREVIOUS DEMO SCREEN
NEXT DEMO SCREEN

FIG. 13A



MAPIT: Claim Selection

Select one of the following claims for viewing and/or clustering by clicking on the claim number.
Preview the application text by clicking on the application number.

Appl/Claim Number	Title
4,775,892	BRIGHTNESS-DEPENDENT FOCUSING AID FOR A MANUALLY FOCUSED VIDE CAMERA
1	1. In video apparatus of the type that isolates the high frequency
2	2. The apparatus as claimed in claim 1 in which said means responsiv
3	3. The apparatus as claimed in claim 2 in which the portion of the
4	4. The apparatus as claimed i claim 1 in which said means responsiv
5	5. The apparatus as claimed in claim 4 in which the control signal i
6	6. The apparatus as claimed in claim 5 in which the d.c. level of th
7	7. In a manually-focused video camera having a signal processing cir
8	8. The apparatus as claimed in claim 7 in which said focus-related
9	9. The apparatus as claimed in claim 7 in which said means for

1312

PREVIOUS DEMO SCREEN
NEXT DEMO SCREEN

FIG. 13B

1320

MAPIT: Claim Viewer

Click on the application number to preview the application text. Click on the "Cluster" button to generate a list of related claims sorted by similarity.

4,775,892	BRIGHTNESS-DEPENDENT FOCUSING AID FOR A MANUALLY FOCUSED VIDEO CAMERA

Cluster Claims

3. The apparatus as claimed in claim 2 in which the portion of the display in which the d.c. level is varied corresponds to a slit-like region of approximately ten video lines.

2. The apparatus as claimed in claim 1 in which said means responsive to variations in the control signal for varying the d.c. level of the video signal is operative during a portion of the display.

1. The video apparatus of the type that isolates the high frequency content of a video signal and uses the high frequency content to modify a video display generated within an electronic viewfinder in order to indicate a properly focused video image, the improvement wherein said video apparatus comprises: means for generating a control signal that varies according to the high frequency content of the video signal as the video image is brought into focus; and means responsive to variations in the control signal for correspondingly varying the d.c. level of the video signal generating the display in the viewfinder whereby the brightness level in the viewfinder corresponds to the high frequency content of the video signal.

PREVIOUS DEMO SCREEN
NEXT DEMO SCREEN

FIG. 13C

MAPIT: Claim Clustering

Results for Application **4,775,892**; Claim 3

4,775,892	BRIGHTNESS-DEPENDENT FOCUSING AID FOR A MANUALLY FOCUSED VIDEO CAMERA
3	3. The apparatus as claimed in claim 2 in which the portion of the ...

Sort by: (●) word vectors ( ) semantic threads

Include text of referred claims? (●) yes ( ) no

Filter out claims from this application? (●) yes ( ) no

Process Results

Click on appl number to view application or claim number to view side-by-side.

FIG. 13D

Rank	Appl/Claim	Description
	[4,794,459]	[COLUMNAR FOCUSING INDICATOR FOR A MANUALLY FOCUSED VID CAMERA]
1.	1	1. In video apparatus of the type that isolates the high frequenc
2.	5	5. The apparatus as claimed in claim 4 in which the portion of t
3.	8	8. The apparatus as claimed in claim 1 in which said control sig
4.	4	4. The apparatus as claimed in claim 1 in which said means res
5.	7	7. The apparatus as claimed in claim 6 in which said gating me
6.	2	2. The apparatus as claimed in claim 1 in which said accumulat
7.	3	3. The apparatus as claimed in claim 2 further including means
8.	10	10. The apparatus as claimed in claim 9 in which said charge st
	[4,660,092]	[FOCUSING AID FOR A MANUALLY FOCUSED VIDEO CAMERA]
9.	1	1. In video apparatus of the type that isolates a focus-related
	[4,794,459]	[COLUMNAR FOCUSING INDICATOR FOR A MANUALLY FOCUSED VID CAMERA]
10.	9	9. In video apparatus of the type that isolates the high-frequenc
	[4,660,092]	[FOCUSING AID FOR A MANUALLY FOCUSED VIDEO CAMERA]
11.	5	5. Apparatus as claimed in claim 4 in which said means for var
	[4,794,459]	[COLUMNAR FOCUSING INDICATOR FOR A MANUALLY FOCUSED VID CAMERA]
12.	6	6. The apparatus as claimed in claim 1 in which said means res
	[4,660,092]	[FOCUSING AID FOR A MANUALLY FOCUSED VIDEO CAMERA]
13.	9	9. Apparatus as claimed in claim 8 in which said means for gen
14.	2	2. The apparatus as claimed in claim 1 in which said means for
15.	4	4. The apparatus as claimed in claim 1 in which said means for
16.	6	6. In a manually-focused video camera having a signal process
17.	10	10. Apparatus as claimed in claim 9 in which said means for gen
18.	3	3. In a video camera having a signal processing circuit of the ty
19.	7	7. Apparatus as claimed in claim 6 in which said circuit path inc
20.	8	8. In a video camera having an image sensor which generates
	[4,794,459]	[COLUMNAR FOCUSING INDICATOR FOR A MANUALLY FOCUSED VID CAMERA]
21.	11	11. In a manually-focused video camera having a signal process
22.	12	12. Apparatus as claimed in claim 11 in which said focus-relatio
23.	13	13. The apparatus as claimed in claim 11 in which said means f
	[4,675,755]	[VIDEO DISK APPARATUS PROVIDING ORGANIZED PICTURE PLAYBA
24.	13	13. Apparatus as claimed in claim 10 wherein said means respo
	[5,034,811]	[VIDEO TRIGGER IN A SOLID STATE MOTION ANALYSIS SYSTEM]
25.	9	9. The system of claims 8 including means for varying the size

1330

4,794,459	COLUMNAR FOCUSING INDICATOR FOR A MANUALLY FOCUSED VIDEO CAMERA
	Recluster on this claim

5. The apparatus as claimed in claim 4 in which the portion of the display in which the d.c. level is subject to change corresponds to a like portion of each line scan and, therefore, to a columnar region of the display.

---

4. The apparatus as claimed in claim 1 in which said means responsive to the amplitude of the accumulated control signal for changing the d.c. level of the video signal is operative during a portion of the display.

---

1. In video apparatus of the type that isolates the hihg frequency content of a video signal and uses the high frewquency content to modify a video display generated within an electronic viewfinder in order to indicate a properly focused video image, the improvement wherein said video apparatus comprises: means for generating a control signal that varies according to the high frequency content of the video signal as the video image is brought into focus; means for accumulating the control signal of the video signal generating the display in the viewfinder whereby the transition in brightness level in the viewfinder corresponds to the high frequency content of the video signal.

Click on "back" button to return to previous demo page.

FIG. 13F

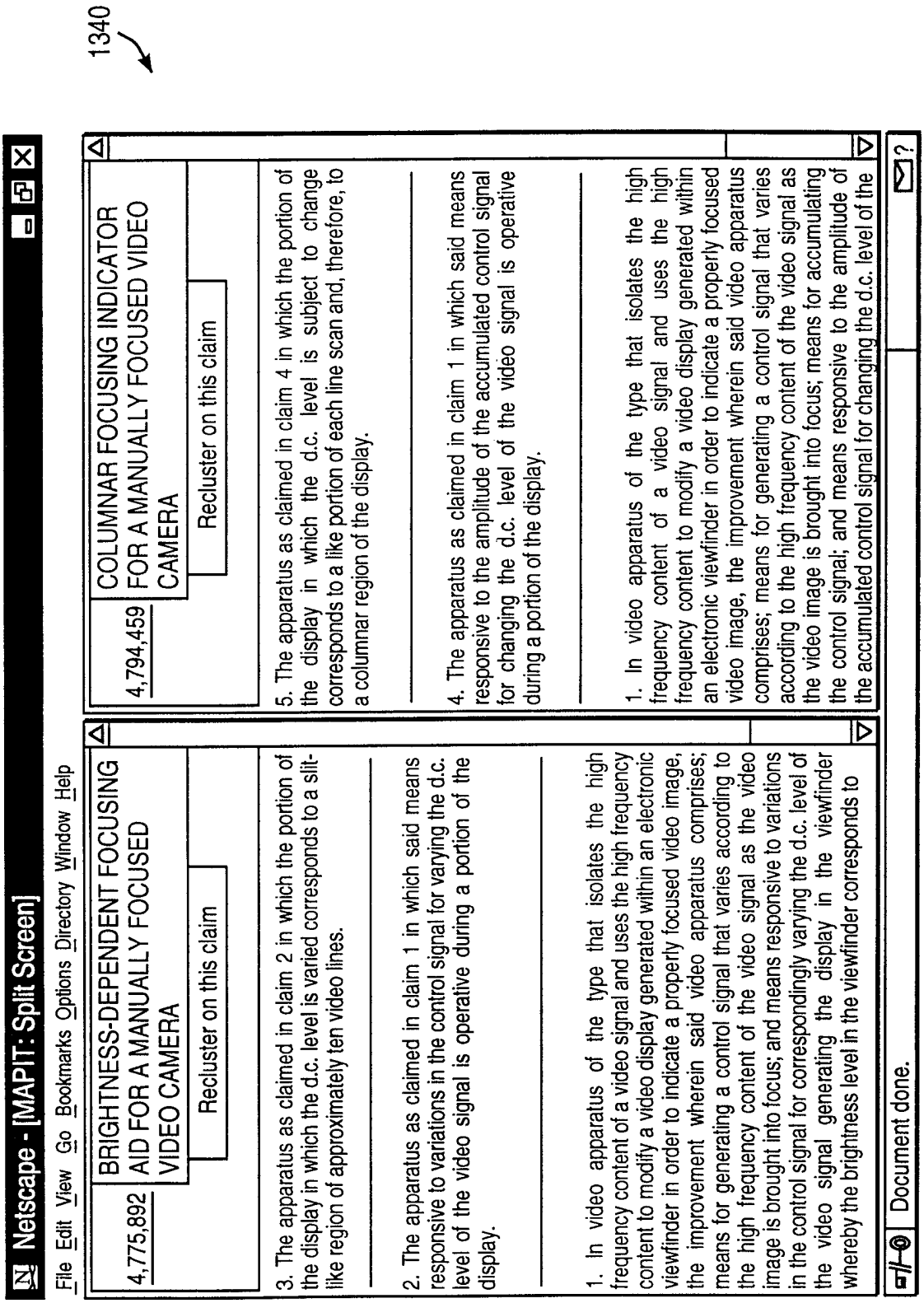

1340
Netscape - [MAPIT: Split Screen]
File Edit View Go Bookmarks Options Directory Window Help
4,775,892
BRIGHTNESS-DEPENDENT FOCUSING AID FOR A MANUALLY FOCUSED VIDEO CAMERA
Recluster on this claim
3. The apparatus as claimed in claim 2 in which the portion of the display in which the d.c. level is varied corresponds to a slit-like region of approximately ten video lines.
2. The apparatus as claimed in claim 1 in which said means responsive to variations in the control signal for varying the d.c. level of the video signal is operative during a portion of the display.
1. In video apparatus of the type that isolates the high frequency content of a video signal and uses the high frequency content to modify a video display generated within an electronic viewfinder in order to indicate a properly focused video image, the improvement wherein said video apparatus comprises; means for generating a control signal that varies according to the high frequency content of the video signal as the video image is brought into focus; and means responsive to variations in the control signal for correspondingly varying the d.c. level of the video signal generating the display in the viewfinder whereby the brightness level in the viewfinder corresponds to
4,794,459
COLUMNAR FOCUSING INDICATOR FOR A MANUALLY FOCUSED VIDEO CAMERA
Recluster on this claim
5. The apparatus as claimed in claim 4 in which the portion of the display in which the d.c. level is subject to change corresponds to a like portion of each line scan and, therefore, to a columnar region of the display.
4. The apparatus as claimed in claim 1 in which said means responsive to the amplitude of the accumulated control signal for changing the d.c. level of the video signal is operative during a portion of the display.
1. In video apparatus of the type that isolates the high frequency content of a video signal and uses the high frequency content to modify a video display generated within an electronic viewfinder in order to indicate a properly focused video image, the improvement wherein said video apparatus comprises; means for generating a control signal that varies according to the high frequency content of the video signal as the video image is brought into focus; means for accumulating the control signal; and means responsive to the amplitude of the accumulated control signal for changing the d.c. level of the
Document done.

FIG. 13G

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/18712

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G06F 17/30
US CL :707/1-6, 10; 345/326,341; 128/653.1
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 707/1-6, 10; 345/326,341; 128/653.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	Please See Continuation of Second Sheet.	

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 07 JANUARY 1998	Date of mailing of the international search report 06 MAR 1998
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer THOMAS G. BLACK Telephone No. (703) 305-9707

Form PCT/ISA/210 (second sheet)(July 1992)★

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/18712

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X,P	US 5,623,681A[RIVETTE et al] 22 April 1997, column 9, lines 51-54, column 10, lines 27-38, column 16, lines 35-37	1,4,6
	column 14, lines 20-25, column 16, lines 39-42	2,5,37
	column 4, lines 23-24	3,8
	column 9, lines 40-46	7
	column 10, lines 58-60	9-12
	column 14, lines 39-44, column 16, lines 39-42	13-18,55
	column 10, lines 19-38, column 16, lines 35-42	19-52
	column 11, lines 44-57	53
	column 12, lines 35-42	54,59-62
	column 13, lines 1-3 and 14-26	56
	column 13, lines 59-60	57
	column 34, lines 37-44	58,63
	column 38, lines 55-67, column 39, line 1-27	64-68