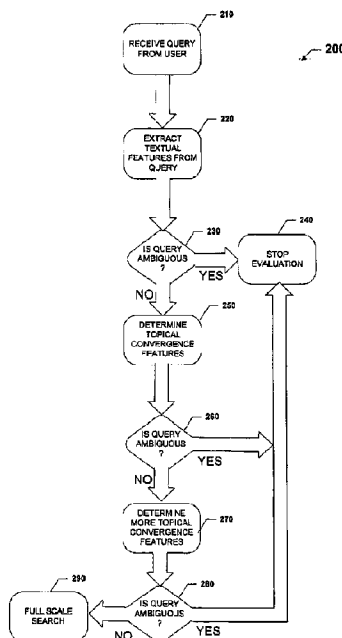




(86) **Date de dépôt PCT/PCT Filing Date:** 2008/12/31
(87) **Date publication PCT/PCT Publication Date:** 2009/07/16
(45) **Date de délivrance/Issue Date:** 2020/03/10
(85) **Entrée phase nationale/National Entry:** 2010/06/30
(86) **N° demande PCT/PCT Application No.:** US 2008/014133
(87) **N° publication PCT/PCT Publication No.:** 2009/088478
(30) **Priorités/Priorities:** 2007/12/31 (US61/009,601);
2007/12/31 (US61/009,684)

(51) **Cl.Int./Int.Cl.** *G06F 16/33* (2019.01)
(72) **Inventeur/Inventor:**
LU, QIANG, US
(73) **Propriétaire/Owner:**
THOMSON REUTERS GLOBAL RESOURCES
UNLIMITED COMPANY, CH
(74) **Agent:** AIRD & MCBURNEY LP

(54) **Titre : SYSTEMES, PROCEDES ET LOGICIEL POUR EVALUER DES INTERROGATIONS D'UTILISATEUR**
(54) **Title: SYSTEMS, METHODS, AND SOFTWARE FOR EVALUATING USER QUERIES**



(57) **Abrégé/Abstract:**

The present inventor devised, among other things, an information retrieval system that determines whether a query is ambiguous or not and based on this determination either continues or aborts a search process. One query evaluation extracts word pairs from an input query and uses features of the extracted word pairs, for example the number of word pairs and their frequencies within a document collection, to determine if the query is ambiguous or not. Another evaluation measures topical convergence, using query related caselaw headnotes that are associated with topics in a legal taxonomy. And yet another checks topical convergence through the lens of full caselaw documents and secondary legal documents, such as law review articles, specifically the minimum number of case law and secondary legal documents that are necessary to span a set of top ranked topics identified using the query.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
16 July 2009 (16.07.2009)

PCT

(10) International Publication Number
WO 2009/088478 A1(51) International Patent Classification:
G06F 17/30 (2006.01)(21) International Application Number:
PCT/US2008/014133(22) International Filing Date:
31 December 2008 (31.12.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/009,684 31 December 2007 (31.12.2007) US
61/009,601 31 December 2007 (31.12.2007) US

(71) Applicant and

(72) Inventor: LU, Qiang [CN/US]; 15 Bayard Court, Pitts-
ford, New York 14534 (US).(74) Agents: STEFFEY, Charles, E. et al.; Schwegman, Lund-
berg & Woessner, P.A., P.O. Box 2938, Minneapolis, Min-
nesota 55402 (US).(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA,
CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE,
EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID,
IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK,
LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW,
MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT,
RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ,
TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM,
ZW.(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM,
ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),
European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,
FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL,
NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG,
CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).**Published:**

— with international search report

(54) Title: SYSTEMS, METHODS AND SOFTWARE FOR EVALUATING USER QUERIES

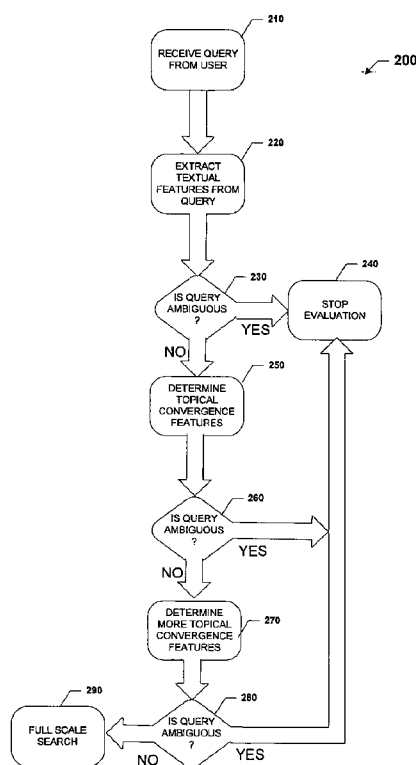


FIGURE 2

(57) Abstract: The present inventor devised, among other things, an information retrieval system that determines whether a query is ambiguous or not and based on this determination either continues or aborts a search process. One query evaluation extracts word pairs from an input query and uses features of the extracted word pairs, for example the number of word pairs and their frequencies within a document collection, to determine if the query is ambiguous or not. Another evaluation measures topical convergence, using query related caselaw headnotes that are associated with topics in a legal taxonomy. And yet another checks topical convergence through the lens of full caselaw documents and secondary legal documents, such as law review articles, specifically the minimum number of case law and secondary legal documents that are necessary to span a set of top ranked topics identified using the query.

WO 2009/088478

PCT/US2008/014133

Systems, Methods, and Software for Evaluating User Queries

5

Copyright Notice and Permission

A portion of this patent document contains material subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent files or records, but otherwise reserves all copyrights whatsoever. The following notice applies to this document: Copyright 2007. Thomson Reuters Global Resources.

10

Related Applications

This application claims priority to U.S. Provisional Application 61/009,684, and also to U.S. Provisional Application 61/009,601, both of which were filed on December 31, 2007.

15

This application is related to U. S. Provisional Application 60/723,322 filed on October 4, 2005 and U.S. application 11/ 538,749 filed on October 4, 2006.

20

Technical Field

Various embodiments of the present invention concern information retrieval systems, particularly online systems that evaluate or screen user queries based on efficacy.

25

Background

The American legal system, as well as some other legal systems around the world, relies heavily on written judicial opinions, the written pronouncements of judges, to articulate or interpret the laws governing resolution of disputes. Each judicial opinion is not only important to resolving a particular legal dispute, but also to resolving similar disputes, or cases, in the future. Because of this, judges and lawyers within our legal system are continually researching an ever-expanding body of past opinions, or case law, for the ones most relevant to resolution of disputes.

30

WO 2009/088478

PCT/US2008/014133

Moreover, judges and lawyers are also regularly researching an expanding body of federal, state, and city-level statutes that articulate specific laws as well as scholarly law review articles that expand on finer legal or philosophical nuances regarding particular legal issues.

- 5 To facilitate these searches Thomson Reuters/West, a division of Thomson Reuters, collects judicial opinions from courts across the United States, and makes them available electronically through its Westlaw™ legal research system. Users access the judicial opinions, for example, by submitting keyword or natural language queries for execution against a jurisdictional database of judicial opinions or case law.
- 10 The Westlaw system also provides access to legal statutes and secondary legal content, such as legal encyclopedia articles and other scholarly material, that are relevant to specific queries.

- A typical user search begins with the user entering and submitting a query: a short sequence of words that ideally defines a specific topic of interest to the user. If
- 15 the query is well formulated or unambiguous, the system returns a set of ranked legal documents that includes the information the user desires within its higher ranks. On the other hand, if the query is ambiguous, the system is vulnerable to returning a set of documents that are not convergent on the specific information needs of the user. Indeed, within the legal domain, the returned documents may span several unrelated
- 20 areas of the law and thus leave the user with a significant task of manually sifting through many irrelevant results.

- Although it is known to alert or warn a user that a given query may yield search results having a large number of documents, for example more than a 1000, there is no indication in this warning that the query itself is ambiguous in terms of the
- 25 diversity of topics that the search results may cover. Moreover, if the query is indeed ambiguous, the search itself is likely to be an inefficient use of both computing resources and professional labor.

- Accordingly, the present inventor recognized a need for better ways of handling the problem of ambiguous queries.

30

Summary

To address this and/or other needs, the present inventor devised, among other things, an information retrieval system that determines whether a query is ambiguous or not and based on this determination either continues or aborts a search process. In the exemplary system, the query evaluation takes at least one of three possible forms. The first form extracts word pairs from an input query and uses features of the extracted word pairs, for example the number of word pairs and their frequencies within a document collection, to determine if the query is ambiguous or not. The second form, which measures topical convergence, uses the query to search a database of caselaw headnotes that are associated with topics in a legal taxonomy. The number of topics implicated by the headnote search is used in combination with one or more of the word pair features to determine if the query is ambiguous or not. The third form also checks topical convergence through the lens of caselaw and secondary legal documents, such as law review articles, specifically the minimum number of case law and secondary legal documents that are necessary to span a set of top ranked topics identified using the query.

Although the exemplary embodiment uses all three forms of ambiguity checks, some embodiments may use only one or two of the ambiguity checks. Also, in some embodiments, an ambiguity check that indicates a query is ambiguous results in the user being prompted for further information to clarify the query.

In accordance with an aspect of the present invention there is provided a method of evaluating ambiguity of a user query, the method comprising:

receiving a user query for searching for responsive documents stored in a first database of documents;

prior to searching the first database, determining a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics that are relevant to the query, determining a set of term frequency inverse

document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;

determining whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and

upon determining that the user query is ambiguous, generating a signal pertaining to the determined ambiguity; or

upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents; and

wherein determining the set of features includes determining a set of caselaw documents based on the query and determining a minimum number of publications required to define a set of publications wherein at least two of the caselaw documents are cited in each of the publications.

In accordance with a further aspect of the present invention there is provided a method of evaluating ambiguity of a user query, the method comprising:

receiving a user query for searching for responsive documents stored in a first database of documents;

prior to searching the first database, determining a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics that are relevant to the query, determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;

determining whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and

upon determining that the user query is ambiguous, generating a signal pertaining to the determined ambiguity; or

upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents; and

wherein the set of derivative values based on the tfidf scores comprise one or more of a sum of tfidf score values, a max value of the tfidf scores, a mean value of the tfidf scores, and a deviation value of the tfidf scores.

In accordance with a further aspect of the present invention there is provided a method of evaluating ambiguity of a user query, the method comprising:

receiving a user query for searching for responsive documents stored in a first database of documents;

prior to searching the first database, determining a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics that are relevant to the query, determining a set of term frequency inverse document frequency ("tfidf") scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;

determining whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and

upon determining that the user query is ambiguous, generating a signal pertaining to the determined ambiguity; or

upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents; and

determining a set of derivative topical divergence feature values;

wherein the determining whether the query is ambiguous or not without searching the first database is based on comparing the tfidf scores to a first constant value, comparing the set of topical divergence features to a second constant value, comparing the derivative values based on the tfidf scores to a third constant value, and comparing the set of derivative topical divergence features to a third constant value.

In accordance with a further aspect of the present invention there is provided an information retrieval system comprising:

- a server having a processor and a memory, further comprising:

- a first database of documents;

- means for receiving by the processor a user query for searching for responsive documents stored in the first database;

- means for determining by the processor prior to searching the first database a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the means for determining the set of features further includes means for determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, means for determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and means for determining a set of derivative values based on the tfidf scores;

- means for determining by the processor whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database;

- means for generating a signal pertaining to query ambiguity upon determining that the user query is ambiguous;

- means for automatically performing a search of the first database for responsive documents upon determining that the query is not ambiguous; and

wherein means for determining the set of features includes means for determining a set of caselaw documents based on the query and means for determining a minimum number of publications required to define a set of publications wherein at least two of the caselaw documents are cited in each of the publications.

In accordance with a further aspect of the present invention there is provided an information retrieval system comprising:

- a server having a processor and a memory, further comprising:

- a first database of documents;

- means for receiving by the processor a user query for searching for responsive documents stored in the first database;

- means for determining by the processor prior to searching the first database a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the means for determining the set of features further includes means for determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, means for determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and means for determining a set of derivative values based on the tfidf scores;

- means for determining by the processor whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database;

- means for generating a signal pertaining to query ambiguity upon determining that the user query is ambiguous;

- means for automatically performing a search of the first database for responsive documents upon determining that the query is not ambiguous; and

wherein the set of derivative values based on the tfidf scores comprise one or more of a sum of tfidf score values, a max value of the tfidf scores, a mean value of the tfidf scores, and a deviation value of the tfidf scores.

In accordance with a further aspect of the present invention there is provided an information retrieval system comprising:

- a server having a processor and a memory, further comprising:

- a first database of documents;

- means for receiving by the processor a user query for searching for responsive documents stored in the first database;

- means for determining by the processor prior to searching the first database a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the means for determining the set of features further includes means for determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, means for determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and means for determining a set of derivative values based on the tfidf scores;

- means for determining by the processor whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database;

- means for generating a signal pertaining to query ambiguity upon determining that the user query is ambiguous;

- means for automatically performing a search of the first database for responsive documents upon determining that the query is not ambiguous;

- means for determining a set of derivative topical divergence feature values; and

wherein the means for determining whether the query is ambiguous or not without searching the first database is based on comparing the tfidf scores to a first constant value, comparing the set of topical divergence features to a second constant value, comparing the derivative values based on the tfidf scores to a third constant value, and comparing the set of derivative topical divergence features to a third constant value.

In accordance with a further aspect of the present invention there is provided a machine-readable medium that stores instructions to be executed by a machine, the instructions when executed causing the machine to:

receive a user query for searching for responsive documents stored in a first database of documents;

prior to searching the first database, determine a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, determining a set of term frequency inverse document frequency ("tfidf") scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;

determine whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and

upon determining that the user query is ambiguous, generate a signal pertaining to the determined ambiguity; or

upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents; and

wherein determining the set of features includes determining a set of caselaw documents based on the query and determining a minimum number of publications

required to define a set of publications wherein at least two of the caselaw documents are cited in each of the publications.

In accordance with a further aspect of the present invention there is provided a machine-readable medium that stores instructions to be executed by a machine, the instructions when executed causing the machine to:

receive a user query for searching for responsive documents stored in a first database of documents;

prior to searching the first database, determine a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;

determine whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and

upon determining that the user query is ambiguous, generate a signal pertaining to the determined ambiguity; or

upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents; and

wherein the set of derivative values based on the tfidf scores comprise one or more of a sum of tfidf score values, a max value of the tfidf scores, a mean value of the tfidf scores, and a deviation value of the tfidf scores.

In accordance with a further aspect of the present invention there is provided a machine-readable medium that stores instructions to be executed by a machine, the instructions when executed causing the machine to:

receive a user query for searching for responsive documents stored in a first database of documents;

prior to searching the first database, determine a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;

determine whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and

upon determining that the user query is ambiguous, generate a signal pertaining to the determined ambiguity; or

upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents;

determining a set of derivative topical divergence feature values; and

wherein the determining whether the query is ambiguous or not without searching the first database is based on comparing the tfidf scores to a first constant value, comparing the set of topical divergence features to a second constant value, comparing the derivative values based on the tfidf scores to a third constant value, and comparing the set of derivative topical divergence features to a third constant value.

Brief Description of Drawings

Figure 1 is a block diagram of an exemplary online legal research system which corresponds to one or more embodiments of the present invention.

Figure 2 is a flowchart of an exemplary method of processing queries which corresponds to one or more embodiments of the present invention.

WO 2009/088478

PCT/US2008/014133

Detailed Description of Exemplary Embodiments

This description, which references and incorporates the above-identified Figures and the appended claims, describes one or more specific embodiments of an invention. These embodiments, offered not to limit but only to exemplify and teach
 5 the invention, are shown and described in sufficient detail to enable those skilled in the art to implement or practice the invention. Thus, where appropriate to avoid obscuring the invention, the description may omit certain information known to those of skill in the art.

10 Exemplary Information-Retrieval System

Figure 1 shows an exemplary online information-retrieval (or legal research) system 100. System 100 includes one or more databases 110, one or more servers 120, and one or more access devices 130.

Databases 110 includes a set of primary databases 112, a set of secondary
 15 databases 114, and a set of metadata databases 116. Primary databases 112, in the exemplary embodiment, include a caselaw database 1121 and a statutes databases 1122, which respectively include judicial opinions and statutes from one or more local, state, federal, and/or international jurisdictions. Secondary databases 114, which contain legal documents of secondary legal authority or more generally
 20 authorities subordinate to those offered by judicial or legislative authority in the primary database, includes an ALR (American Law Reports) database, 1141, an AMJUR database 1142, a West Key Number (KNUM) Classification database 1143, and a law review (LREV) database 1144, as well as other databases that may be mentioned below. Metadata databases 116 includes case law and statutory citation
 25 relationships, KeyCite data (depth of treatment data, quotation data, headnote assignment data, and ResultsPlus secondary source recommendation data. Also, in some embodiments, primary and secondary connote the order of presentation of search results and not necessarily the authority or credibility of the search results.

Databases 110, which take the exemplary form of one or more electronic,
 30 magnetic, or optical data-storage devices, include or are otherwise associated with respective indices (not shown). Each of the indices includes terms and phrases in association with corresponding document addresses, identifiers, and other conventional information. Databases 110 are coupled or couplable via a wireless or

WO 2009/088478

PCT/US2008/014133

wireline communications network, such as a local-, wide-, private-, or virtual-private network, to server 120.

Server 120, which is generally representative of one or more servers for serving data in the form of webpages or other markup language forms with associated applets, ActiveX controls, remote-invocation objects, or other related software and data structures to service clients of various "thicknesses." More particularly, server 120 includes a processor module 121, a memory module 122, a subscriber database 123, a primary search module 124, query discriminator module 125, and a user-interface module 126.

Processor module 121 includes one or more local or distributed processors, controllers, or virtual machines. In the exemplary embodiment, processor module 121 assumes any convenient or desirable form.

Memory module 122, which takes the exemplary form of one or more electronic, magnetic, or optical data-storage devices, stores subscriber database 123, primary search module 124, query discriminator (or evaluation) module 125, and user-interface module 126.

Subscriber database 123 includes subscriber-related data for controlling, administering, and managing pay-as-you-go or subscription-based access of databases 110. In the exemplary embodiment, subscriber database 123 includes one or more preference data structures.

Primary search module 124 includes one or more search engines and related user-interface components, for receiving and processing user queries against one or more of databases 110. In the exemplary embodiment, one or more search engines associated with search module 124 provide Boolean, tf-idf, natural-language search capabilities. Additionally, module 124 includes search capabilities described in co-pending application 11/ 538,749 which is referenced above as well as Results Plus search capabilities described in co-pending U.S. patent application 11/028,476 which was filed on January 3, 2005.

Query discriminator module 125 includes machine-executable instructions and related machine-readable data, such as computed features for evaluating the efficacy or ambiguity of queries as further described below. In the exemplary embodiment, module 125 includes one or more feature vector builders and learning machines to implement the functionality described herein.

WO 2009/088478

PCT/US2008/014133

User-interface module 126 includes machine readable and/or executable instruction sets for wholly or partly defining web-based user interfaces, such as search interface 1261 and results interface 1262, over a wireless or wireline communications network on one or more access devices, such as access device 130.

5 Access device 130 is generally representative of one or more access devices. In the exemplary embodiment, access device 130 takes the form of a personal computer, workstation, personal digital assistant, mobile telephone, or any other device capable of providing an effective user interface with a server or database. Specifically, access device 130 includes a processor module 131 one or more
10 processors (or processing circuits) 131, a memory 132, a display 133, a keyboard 134, and a graphical pointer or selector 135.

Processor module 131 includes one or more processors, processing circuits, or controllers. In the exemplary embodiment, processor module 131 takes any convenient or desirable form. Coupled to processor module 131 is memory 132.

15 Memory 132 stores code (machine-readable or executable instructions) for an operating system 136, a browser 137, and a graphical user interface (GUI) 138. In the exemplary embodiment, operating system 136 takes the form of a version of the Microsoft Windows operating system, and browser 137 takes the form of a version of Microsoft Internet Explorer. Operating system 136 and browser 137 not only receive
20 inputs from keyboard 134 and selector 135, but also support rendering of GUI 138 on display 133. Upon rendering, GUI 138 presents data in association with one or more interactive control features (or user-interface elements). (The exemplary embodiment defines one or more portions of interface 138 using applets or other programmatic objects or structures from server 120.)

25 More specifically, graphical user interface 138 defines or provides one or more display regions, such as a query or search region 1381 and a search-results region 1382. Query region 1381 is defined in memory and upon rendering includes one or more interactive control features (elements or widgets), such as a query input region 1381A, a query submission button 1381B. Search-results region 1382 is also
30 defined in memory and upon rendering presents a variety of types of information in response to a query submitted in region 1381. In the exemplary embodiment, the results region identifies one or more source case law documents (that is, one or more good cases, usually no more than five), jurisdictional information, issues information, additional key cases, key statutes, key briefs or trial documents, key analytical

WO 2009/088478

PCT/US2008/014133

materials, and/or additional related materials. (See Figure 3, which is described below, for a more specific example of a results region.) Each identified document in region 1382 is associated with one or more interactive control features, such as hyperlinks, not shown here. User selection of one or more of these control features results in retrieval and display of at least a portion of the corresponding document within a region of interface 138 (not shown in this figure.) Although Figure 1 shows query region 1381 and results region 1382 as being simultaneously displayed, some embodiments present them at separate times.

10 Exemplary Method of Screening or Processing Queries

Figure 2 shows a flowchart 200 of an exemplary method of evaluating the quality of a user query. The flowchart includes process blocks 210-XXX.

In block 210, the exemplary method begins with receipt of a user query. In the exemplary embodiment, receiving a user query entails a user logging into an online legal research system, such as the Westlaw system, using a client access device, and entering and submitting a query through a graphical user interface into the system. Submission of the query results in communication of the query over a wide- or local-area network to a server within the system, with the server including a processor and memory storing machine executable instructions for processing a query as described herein. Execution continues at block 220.

Block 220 entails extracting textual features from the query. In the exemplary embodiment, this extraction entails removing stop words, punctuation marks, and other symbols from the query. It also entails removing person names, with this removal based on a name list. Exemplary stop words include common words, such as "the," "a," "and," "who," "for" and "is." The person names query removes person names from the query. (The exemplary embodiment, which is based on a legal domain, assumes that if a query only contains person names, the user is most likely trying to retrieve a specific case; therefore the query is not a good query.) Exceptions exist; for example, the query "Heck Balisok" is considered good since it involves two different case names and the user is looking for the issues in these cases, not the cases themselves.

After stripping the unwanted text, the exemplary embodiment groups the remaining terms into a set of one or more word pairs. The table below shows two

WO 2009/088478

PCT/US2008/014133

sample queries and the extracted word pairs. (Note that some embodiments may not employ stemming as indicated in the table.)

Sample Query	Extracted Word Pairs
Who is a public figure for speech defamation?	1) public-public, 2) public-figure, 3) public-4) speech, 5) public-defamati, 6) figure-figure, 7) speech-figure, 8) defamati-figure, 9) speech-speech, 10) speech-defamati, 11) defamati-defamati.
Berderka misrepresentations unreported decisions	1) decision-decision; 2) unreport-unreport; 3) misrepre-misrepre; 4) decision-misrepre, 5) decision-unreport

- 5 After the word pairs are defined the exemplary embodiment determines the following set of query features:
1. Number of Word pairs. If a query contains text with legal concepts, it is likely to be a good query.
 2. tfidf score for each word pair. a tfidf score is determined for each word pair based on a dictionary of word pairs that are extracted from a corpus of legal headnotes. (Headnotes are understood to be brief statements regarding the legal holdings and principles from judicial opinions.) The word pairs are defined as words within a certain distance, for example five words, of each other within the query.
 3. Sum of tfidf Scores values for the set of Word pairs. The sum of tfidf values of all the Word pairs. (tfidf is the product of the term frequency (tf) and the inverse document frequency (idf) of the term.)
 4. Max value of tfidf Score for Word pairs. The maximal tfidf value of all the Word pairs.
 5. Mean tfidf Score for Word pairs. The mean tfidf score is computed as the sum of valid word pairs' tfidf score normalized or divided by the number of valid word pairs.
 6. Deviation tfidf Score for Word pairs. The deviation of the tfidf values of all the Word pairs.
- 25 Execution then proceeds to block 230.

WO 2009/088478

PCT/US2008/014133

Block 230 entails determine whether the query is bad, in other words ambiguous. In the exemplary embodiment, this entails applying the following set of rules to one or more of the extracted features:

- a. if (is single word query) and (number of word pairs) ≤ 2 , then designate query as bad;
- b. if (number of word pairs) ≤ 5 , and if $9 < (\text{sum of TFIDF}) \leq 14$, and if (max value of TFIDF > 4.6), then designate query as bad;
- c. if neither of conditions a or b is true, continue evaluation of query.

If the query is determined as bad, that is ambiguous, execution branches to block 240 which terminates the query processing in the exemplary embodiment. However, some embodiments alert the user that the query has been deemed ambiguous and request further information or query revision from the user. If further information, such as revised query is submitted, execution returns to block 220 for extraction of features from the revised query. On the other hand, if the determination is not bad, further evaluation of the query continues at block 250. (Some embodiments may skip this further evaluation and proceed to full scale searching based on the query.)

Block 250 entails extracting or determining topical divergence features regarding the query. In the exemplary embodiment, this entails determining a number of legal topical or subject matter categories from a topical hierarchy that are related to the query as well as a number of secondary law documents related to the query. To this end, the exemplary embodiment conducts two separate searches based on the query: a search of a headnote database for the top 25, 50, 75, or 100 headnotes and their associated KeyNumberTM topical codes, and a search of a secondary legal materials database to determine the number of relevant secondary documents. The KeyNumber topic codes are then clustered, with the number of resulting clusters defined as the headnote topic number for use below. The secondary legal materials search is conducted using the WestlawTM ResultsPlusTM document suggestion engine, which is described in copending U.S. patent application 11/028,476 which was filed on January 3, 2005.

The suggestion engine searches for documents from the following secondary materials: American Law Reports (ALR), American Jurisprudence (AMJ), TRE, and LWR and the number of returned documents is designated as "number of RP suggestions." The headnote topic number and number of RP suggestions are used in block 260.

WO 2009/088478

PCT/US2008/014133

Block 260 entails determining or classifying whether the query is bad (ambiguous) based on the topical divergence features. In making this determination, the exemplary embodiment uses the following rules:

- a. if (deviation of TFIDF) $\leq$ 0.63 and (Headnote topic number) $\leq$ 0, then
5 designate the query as bad;
- b. if (deviation of TFIDF $\leq$ 0.63) and (Headnote topic number) $>$ 15, then
designate the query as bad;
- c. if (deviation of TFIDF $\leq$ 0.63) and (headnote topic number) $>$ 8 and (Headnote
topic number) $\leq$ 15 and (number of word pairs) $\leq$ 3 and (number of RP
10 suggestions) $>$ 11, then designate the query as bad;
- d. if (deviation of TFIDF) $>$ 0.63 and (number of word pairs) $\leq$ 5 and (Headnote
topic number) $>$ 6 and (number of word pairs) $\leq$ 2 and (number of PR
suggestions) $>$ 14, then designate the query as bad;
- e. if (deviation of TFIDF) $>$ 0.63 and (number of word pairs) $\leq$ 5 and (Headnote
15 topic numbers $>$ 6 and (number of word pairs) $>$ 2 and (mean of TFIDF) $>$ 3.37
and (maximal value of TFIDF) $\leq$ 5.86), then designate the query as bad;
- f. if neither of conditions a, b, c, d, or e is true, continue evaluation of query.

If the query is determined as bad, that is ambiguous, execution branches to block 240 which terminates the query processing in the exemplary embodiment. However,
20 some embodiments alert the user that the query has been deemed ambiguous and request further information or query revision from the user. If further information, such as revised query is submitted, execution returns to block 220 for extraction of features from the revised query. On the other hand, if all the above conditions are false, meaning the query was not designated as bad, further evaluation of the query
25 continues at block 270. (Some embodiments may skip this further evaluation and proceed to full scale searching based on the query.)

Block 270 entails extracting additional process or topical divergence features related to the query. In the exemplary embodiment, this extraction of process features is a two-part process. The first part entails use of a KeyNumber topic finder, and a
30 case-finder. The KeyNumber topic finder takes the query text as input and outputs a number of West Key Numbers topic codes (for example ten to twenty five of them) according to the relevance to the search text as was done at block 240. The exemplary embodiment then clusters these top ranked KeyNumbers, for example, the top 15, to determine a number of relevant KeyNumber topics.

WO 2009/088478

PCT/US2008/014133

To find or identify relevant caselaw, the exemplary embodiment takes the user's query and 15 top ranked Key Numbers topical codes as inputs and retrieves a set of one or more caselaw documents based on their relevance. It then measures the topical convergence of these caselaw documents. With caselaw documents, however,
 5 measuring topical convergence by clustering them into digest topics is less useful as an indicator than it was with headnotes (as done at block 240) since one caselaw document may discuss various legal topics. So rather than rely on digest topics to determine a measure of topical convergence, the exemplary embodiment uses citation information.

Specifically, the exemplary embodiment uses Westlaw™ KeyCite™ citation table
 10 to determine the minimum number of publications that need to be collected to cite all top N ranked cases (for example N=10 or N=15) cases that were retrieved based on KeyNumber topic. This minimum number of publications necessary to "span" the list of cases by virtue of a co-citations is indicative of the topical convergence of the retrieved documents and thus the relative ambiguity of the query. In other words, the top 10 cases
 15 are searched against the citation table to see how many of them are co-cited in the same publication (either a secondary law document or a case). The idea is that if two cases are co-cited in the same publication, they probably deal with the same or similar legal issues. The minimal number of publications which cover all the top 10 cases is designated as feature QD. Execution continues at block 280.

20 Block 280 entails determining or classifying whether the query is bad (ambiguous) based on the topical divergence features. In making this determination, the exemplary embodiment uses the following rules:

- a. if (number of word pairs) $\leq$ 2 and (Key number Finder (KNF) topic) $>$ 12, then designate query as bad;
- 25 b. if (deviation of TFIDF) $>$ 0.63 and (number of word pairs) $\leq$ 5 and (Headnote search topics) $>$ 6 and (number of RP suggestions) $>$ 6 and (KNF topic) $>$ 11, then designate query as bad;
- c. if (number of word pairs) $\leq$ 2 and (KNF topic) $>$ 9 and (Headnote search topic) $>$ 15, then designate query as bad;
- 30 d. if (number of word pairs) $\leq$ 5 and (Headnote search topic) $>$ 12 and (KNF topic) $>$ 9 and (top document cites number of top 10 cases) $\leq$ 3 and (Top 10 cases cited by number of documents) $>$ 6 and (Top 5 documents cite number of top 10 cases in percentage) $\leq$ 0.4, then designate query as bad;

WO 2009/088478

PCT/US2008/014133

- e. if (number of word pairs) >2 and (number of word pairs) ≤ 5 and (Headnote search topics) >7 and (KNF topic >11) and (Top 5 documents cite number of top 10 cases in percentage) ≤ 0.4 and (Top 10 cases cited by number of documents) >7), then designate query as bad;
 - 5 f. if (deviation of TFIDF) ≤ 0.63 and (KNF topic) >9 and ((top document cites number of top 10 cases) ≤ 4 , then designate query as bad;
 - g. if neither of conditions a, b, c, d, e, or f is true, continue evaluation of query.
- If the query is determined as bad, that is ambiguous, execution branches to block 240 which terminates the query processing in the exemplary embodiment. However,
- 10 some embodiments alert the user that the query has been deemed ambiguous and request further information or query revision from the user. If further information, such as revised query is submitted, execution returns to block 220 for extraction of features from the revised query or advanced directly to full scale searching at block 290. On the other hand, if all the above conditions are false, meaning the query was
- 15 not designated as bad, further evaluation of the query continues at block 290.

Other Embodiments

Some embodiments may use the features noted above in more complex form of good-bad ambiguity classifiers rather than the heuristic ones of the exemplary

20 embodiment. For example, machine learning algorithms, decision rules, decision trees, neural networks, statistical classifiers, binary classifiers (support vector machines, for example), other learning algorithms, or the like may be used to classify queries as ambiguous or not. For example, a set of decision rules may be trained using training examples with known clarity qualities to classify & a query. The input

25 of the decision rules is the set of features. The decision rules are then trained to classify queries into bad or not bad.

Additional embodiments of the present invention include presenting the user with

options if the query is determined as bad. One embodiment tells the user that the

30 query was insufficient to run a complete search report. Another embodiment detects an ambiguous query and responds by prompting the user to be more specific about the query terms. Yet another embodiment could present the user with a partial search report in hopes that this will aid the user in clarifying the query.

WO 2009/088478

PCT/US2008/014133

Conclusion

The embodiments described above are intended only to illustrate and teach one or more ways of practicing or implementing the present invention, not to restrict its breadth or scope. The actual scope of the invention, which embraces all ways of practicing or implementing the teachings of the invention, is defined only by the following claims and their equivalents.

What is claimed is:

1. A method of evaluating ambiguity of a user query, the method comprising:
 - receiving a user query for searching for responsive documents stored in a first database of documents;
 - prior to searching the first database, determining a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics that are relevant to the query, determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;
 - determining whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and
 - upon determining that the user query is ambiguous, generating a signal pertaining to the determined ambiguity; or
 - upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents; and
 - wherein determining the set of features includes determining a set of caselaw documents based on the query and determining a minimum number of publications required to define a set of publications wherein at least two of the caselaw documents are cited in each of the publications.
2. The method of claim 1, wherein determining whether the query is ambiguous includes determining a value based on frequency of one or more of the word pairs within a collection of documents.

3. The method of claim 1, further comprising outputting, in response to a determination that the query is ambiguous, a message to a client access device, requesting that a user submit further information regarding the query.
4. The method of claim 1, further comprising searching the first database based on the query, in response to the determination that the query is not ambiguous.
5. A method of evaluating ambiguity of a user query, the method comprising:
 - receiving a user query for searching for responsive documents stored in a first database of documents;
 - prior to searching the first database, determining a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics that are relevant to the query, determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;
 - determining whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and
 - upon determining that the user query is ambiguous, generating a signal pertaining to the determined ambiguity; or
 - upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents; and

wherein the set of derivative values based on the tfidf scores comprise one or more of a sum of tfidf score values, a max value of the tfidf scores, a mean value of the tfidf scores, and a deviation value of the tfidf scores.

6. A method of evaluating ambiguity of a user query, the method comprising:

receiving a user query for searching for responsive documents stored in a first database of documents;

prior to searching the first database, determining a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics that are relevant to the query, determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;

determining whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and

upon determining that the user query is ambiguous, generating a signal pertaining to the determined ambiguity; or

upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents; and

determining a set of derivative topical divergence feature values;

wherein the determining whether the query is ambiguous or not without searching the first database is based on comparing the tfidf scores to a first constant value, comparing the set of topical divergence features to a second constant value, comparing the derivative values based on the tfidf scores to a third constant value, and comparing the set of derivative topical divergence features to a third constant value.

7. An information retrieval system comprising:
- a server having a processor and a memory, further comprising:
 - a first database of documents;
 - means for receiving by the processor a user query for searching for responsive documents stored in the first database;
 - means for determining by the processor prior to searching the first database a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the means for determining the set of features further includes means for determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, means for determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and means for determining a set of derivative values based on the tfidf scores;
 - means for determining by the processor whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database;
 - means for generating a signal pertaining to query ambiguity upon determining that the user query is ambiguous;
 - means for automatically performing a search of the first database for responsive documents upon determining that the query is not ambiguous; and
 - wherein means for determining the set of features includes means for determining a set of caselaw documents based on the query and means for determining a minimum number of publications required to define a set of publications wherein at least two of the caselaw documents are cited in each of the publications.

8. The system of claim 7, wherein the means for determining whether the query is ambiguous includes means for determining a value based on frequency of one or more of the word pairs within a collection of documents outside the first database.
9. The system of claim 7, further comprising means for outputting, in response to a determination that the query is ambiguous, a message to a client access device, requesting that a user submit further information regarding the query.
10. The system of claim 7, further comprising means for searching the first database based on the query, in response to the determination that the query is not ambiguous.
11. An information retrieval system comprising:
 - a server having a processor and a memory, further comprising:
 - a first database of documents;
 - means for receiving by the processor a user query for searching for responsive documents stored in the first database;
 - means for determining by the processor prior to searching the first database a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the means for determining the set of features further includes means for determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, means for determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and means for determining a set of derivative values based on the tfidf scores;

means for determining by the processor whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database;

means for generating a signal pertaining to query ambiguity upon determining that the user query is ambiguous;

means for automatically performing a search of the first database for responsive documents upon determining that the query is not ambiguous; and

wherein the set of derivative values based on the tfidf scores comprise one or more of a sum of tfidf score values, a max value of the tfidf scores, a mean value of the tfidf scores, and a deviation value of the tfidf scores.

12. An information retrieval system comprising:

a server having a processor and a memory, further comprising:

a first database of documents;

means for receiving by the processor a user query for searching for responsive documents stored in the first database;

means for determining by the processor prior to searching the first database a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the means for determining the set of features further includes means for determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, means for determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and means for determining a set of derivative values based on the tfidf scores;

means for determining by the processor whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database;

means for generating a signal pertaining to query ambiguity upon determining that the user query is ambiguous;

means for automatically performing a search of the first database for responsive documents upon determining that the query is not ambiguous;

means for determining a set of derivative topical divergence feature values; and

wherein the means for determining whether the query is ambiguous or not without

searching the first database is based on comparing the tfidf scores to a first constant value, comparing the set of topical divergence features to a second constant value, comparing the derivative values based on the tfidf scores to a third constant value, and comparing the set of derivative topical divergence features to a third constant value.

13. A machine-readable medium that stores instructions to be executed by a machine, the instructions when executed causing the machine to:

receive a user query for searching for responsive documents stored in a first database of documents;

prior to searching the first database, determine a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;

determine whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and
upon determining that the user query is ambiguous, generate a signal pertaining to the determined ambiguity; or
upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents; and

wherein determining the set of features includes determining a set of caselaw documents based on the query and determining a minimum number of publications required to define a set of publications wherein at least two of the caselaw documents are cited in each of the publications.

14. The machine-readable medium of claim 13, wherein the determining whether the query is ambiguous includes determining a value based on frequency of one or more of the word pairs within a collection of documents outside the first database.
15. The machine-readable medium of claim 13, further comprising outputting, in response to a determination that the query is ambiguous, a message to a client access device, requesting that a user submit further information regarding the query.
16. A machine-readable medium that stores instructions to be executed by a machine, the instructions when executed causing the machine to:
 - receive a user query for searching for responsive documents stored in a first database of documents;
 - prior to searching the first database, determine a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are

relevant to the query, determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;

determine whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and

upon determining that the user query is ambiguous, generate a signal pertaining to the determined ambiguity; or

upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents; and

wherein the set of derivative values based on the tfidf scores comprise one or more of a sum of tfidf score values, a max value of the tfidf scores, a mean value of the tfidf scores, and a deviation value of the tfidf scores.

17. A machine-readable medium that stores instructions to be executed by a machine, the instructions when executed causing the machine to:

receive a user query for searching for responsive documents stored in a first database of documents;

prior to searching the first database, determine a set of features based on the user query including a first step of extracting textual information from the query to form extracted features, the extracted features including at least one from the set consisting of removing stop words, removing personal names, removing punctuation marks, and removing non-alphabetical symbols, and a second step of extracting a set of word pairs from the query, wherein the determining the set of features further includes determining a number of topics in a taxonomy that are relevant to the query, determining a set of topical divergence features based on the determined number of topics in the taxonomy that are relevant to the query, determining a set of term frequency inverse document frequency (“tfidf”) scores for each word pair, the tfidf scores comprising a product of the term frequency and the inverse document frequency for each word pair, and determining a set of derivative values based on the tfidf scores;

determine whether the query is ambiguous or not based on the set of features and the set of topical divergence features without searching the first database; and
upon determining that the user query is ambiguous, generate a signal pertaining to the determined ambiguity; or
upon determining that the query is not ambiguous, automatically performing a search of the first database for responsive documents;
determining a set of derivative topical divergence feature values; and
wherein the determining whether the query is ambiguous or not without searching the first database is based on comparing the tfidf scores to a first constant value, comparing the set of topical divergence features to a second constant value, comparing the derivative values based on the tfidf scores to a third constant value, and comparing the set of derivative topical divergence features to a third constant value.

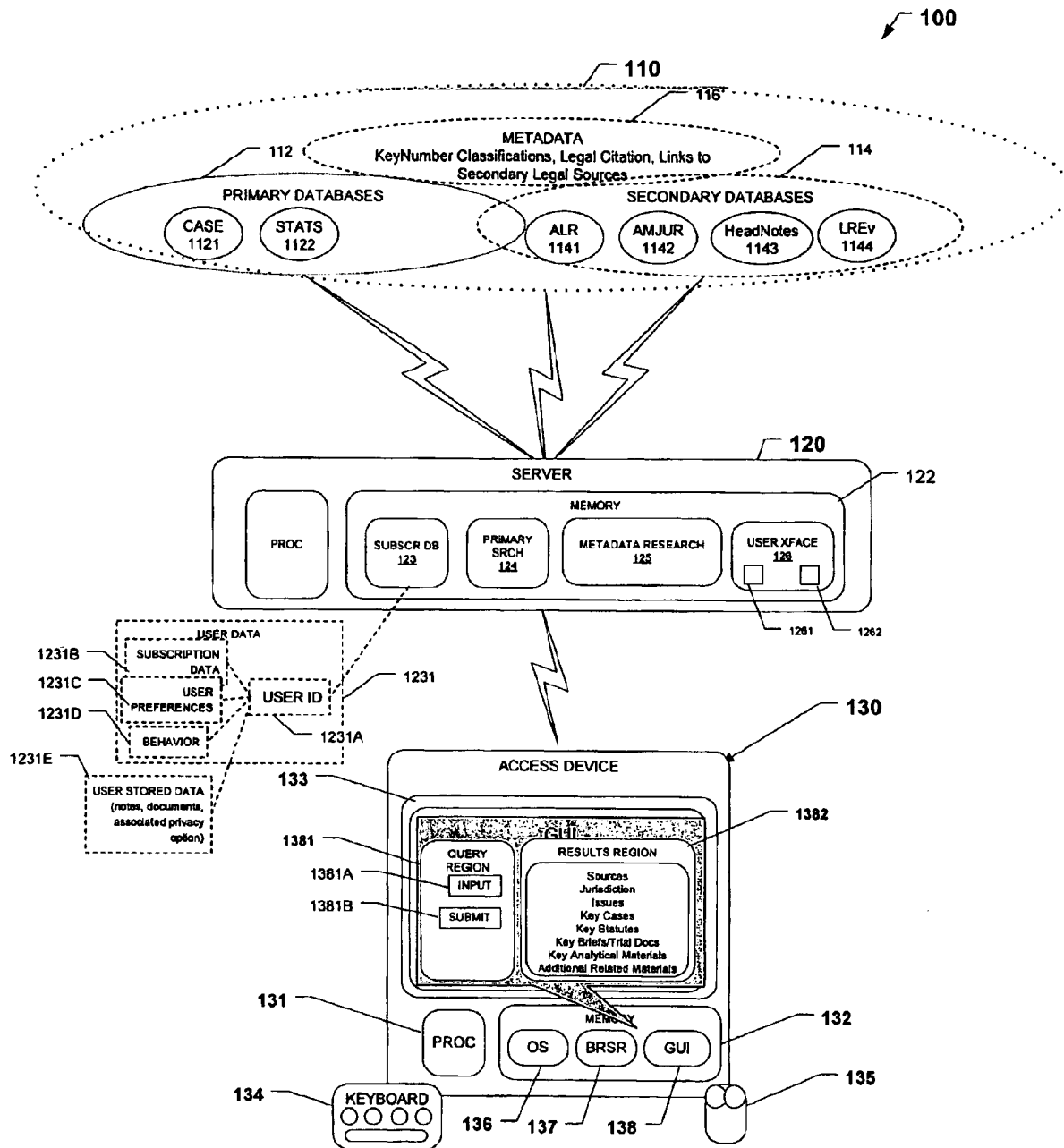


FIGURE 1

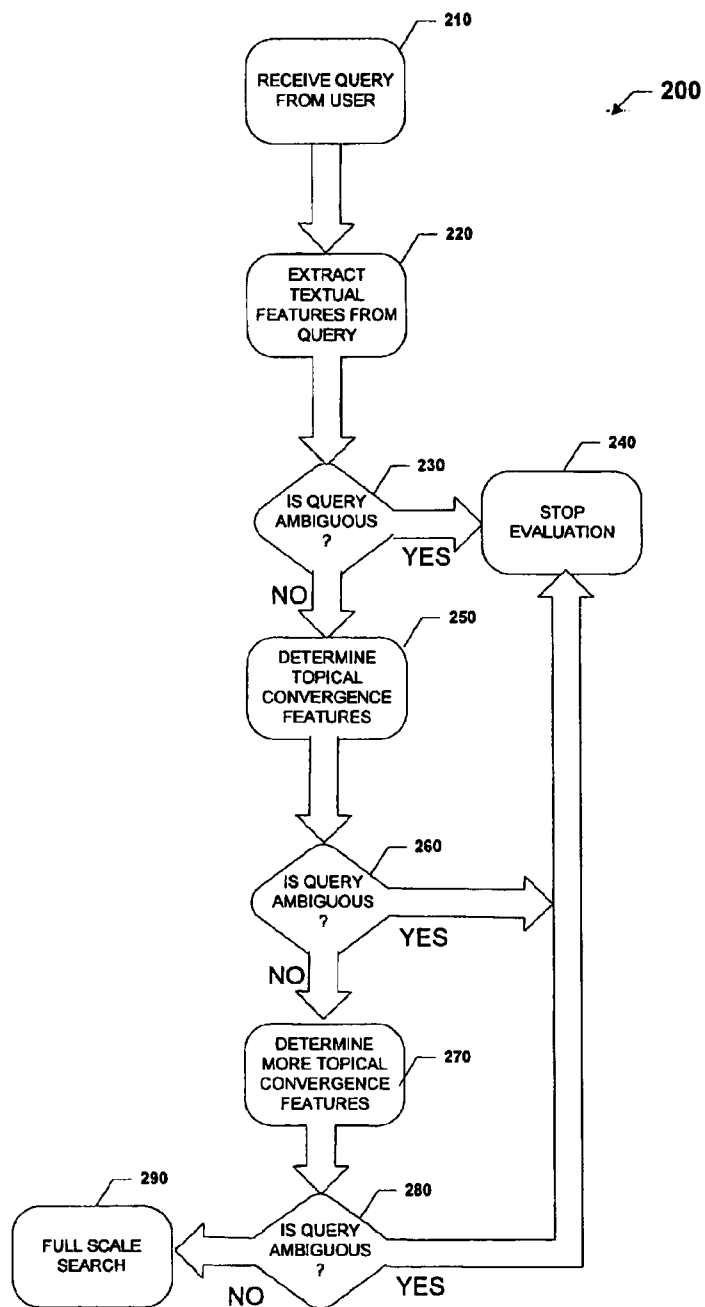


FIGURE 2

