Xerox Corp. v. Google, Inc.
Case No. 10-136-LPS

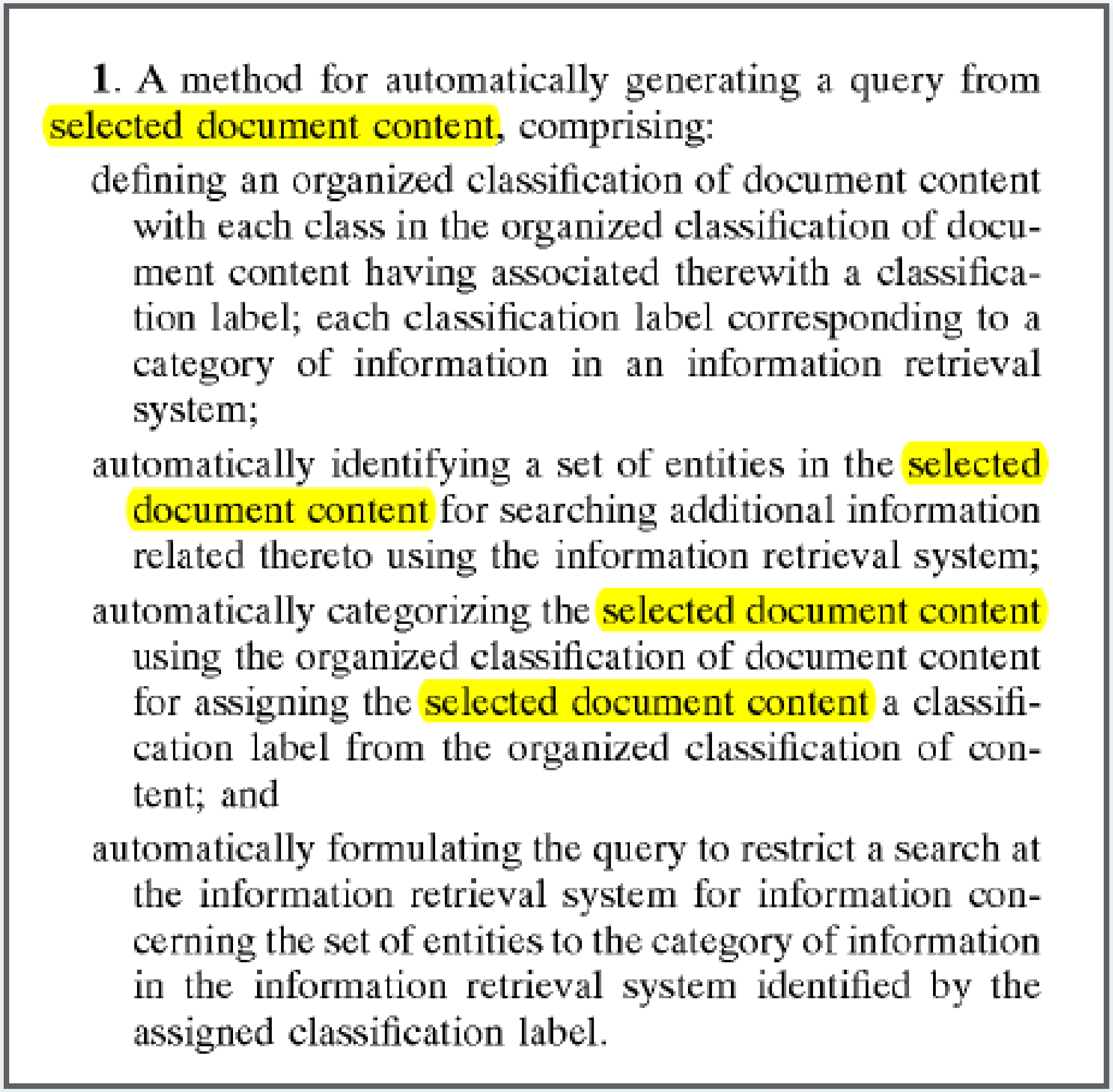
xerox

Claim Construction Hearing for

U.S. Patent 6,778,979

May 19, 2011

“selected document content”



1. A method for automatically generating a query from selected document content, comprising:

defining an organized classification of document content with each class in the organized classification of document content having associated therewith a classification label; each classification label corresponding to a category of information in an information retrieval system;

automatically identifying a set of entities in the selected document content for searching additional information related thereto using the information retrieval system;

automatically categorizing the selected document content using the organized classification of document content for assigning the selected document content a classification label from the organized classification of content; and

automatically formulating the query to restrict a search at the information retrieval system for information concerning the set of entities to the category of information in the information retrieval system identified by the assigned classification label.

“selected document content”

Parties’ Constructions

<u>Xerox</u>	<u>Defendants</u>
“all or part of the content of a document in electronic form”	Indefinite

Law on Indefiniteness

> “A claim will be found indefinite only if it ‘is insolubly ambiguous, and no narrowing construction can properly be adopted . . . .’”

Praxair, Inc. v. ATMI, Inc., 543 F.3d 1306, 1319 (Fed. Cir. 2008);
Leader Techs., Inc. v. Facebook, Inc., 692 F.Supp.2d 425, 436
(D. Del. 2010).

“selected document content”

“Selected document content” is the input to the claimed method (*i.e.*, the document content in which entities are identified and that is categorized).

selected document content

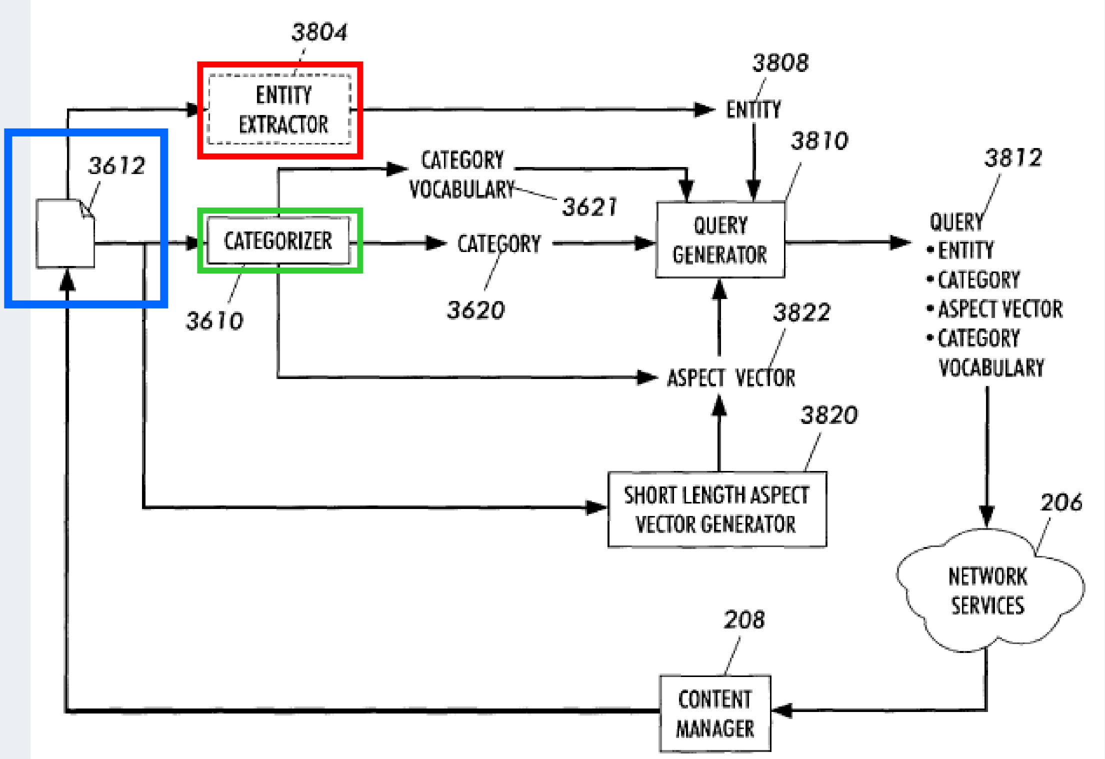
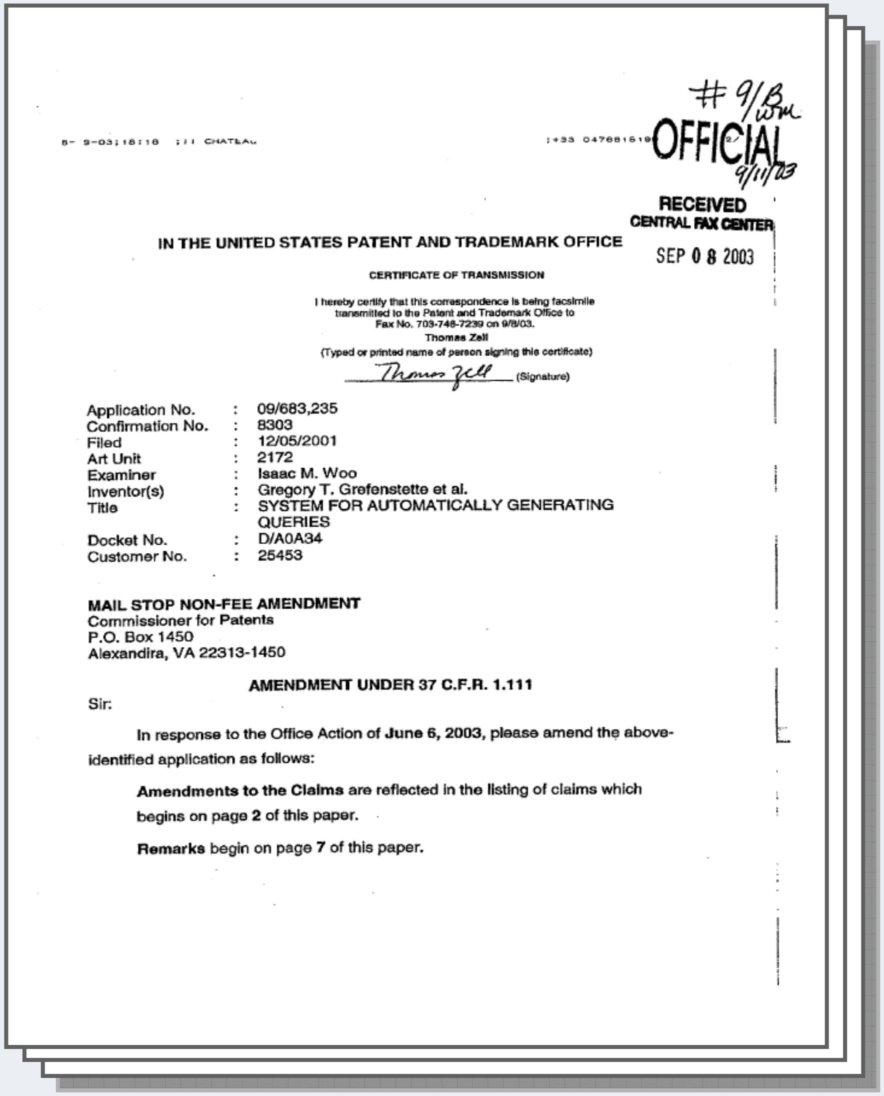


FIG. 38

“selected document content”



8- 9-03;18:18 ;11 CHATEAU ;+33 0476815191 2/

9/8 wm

OFFICIAL 9/11/03

RECEIVED
CENTRAL FAX CENTER

SEP 0 8 2003

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

CERTIFICATE OF TRANSMISSION

I hereby certify that this correspondence is being facsimile
transmitted to the Patent and Trademark Office to
Fax No. 703-746-7239 on 9/8/03.

Thomas Zell
(Typed or printed name of person signing this certificate)

Thomas Zell (Signature)

Application No.	:	09/683,235
Confirmation No.	:	8303
Filed	:	12/05/2001
Art Unit	:	2172
Examiner	:	Isaac M. Woo
Inventor(s)	:	Gregory T. Grefenstette et al.
Title	:	SYSTEM FOR AUTOMATICALLY GENERATING QUERIES
Docket No.	:	D/A0A34
Customer No.	:	25453

MAIL STOP NON-FEE AMENDMENT
Commissioner for Patents
P.O. Box 1450
Alexandira, VA 22313-1450

AMENDMENT UNDER 37 C.F.R. 1.111

Sir:

In response to the Office Action of **June 6, 2003**, please amend the above-identified application as follows:

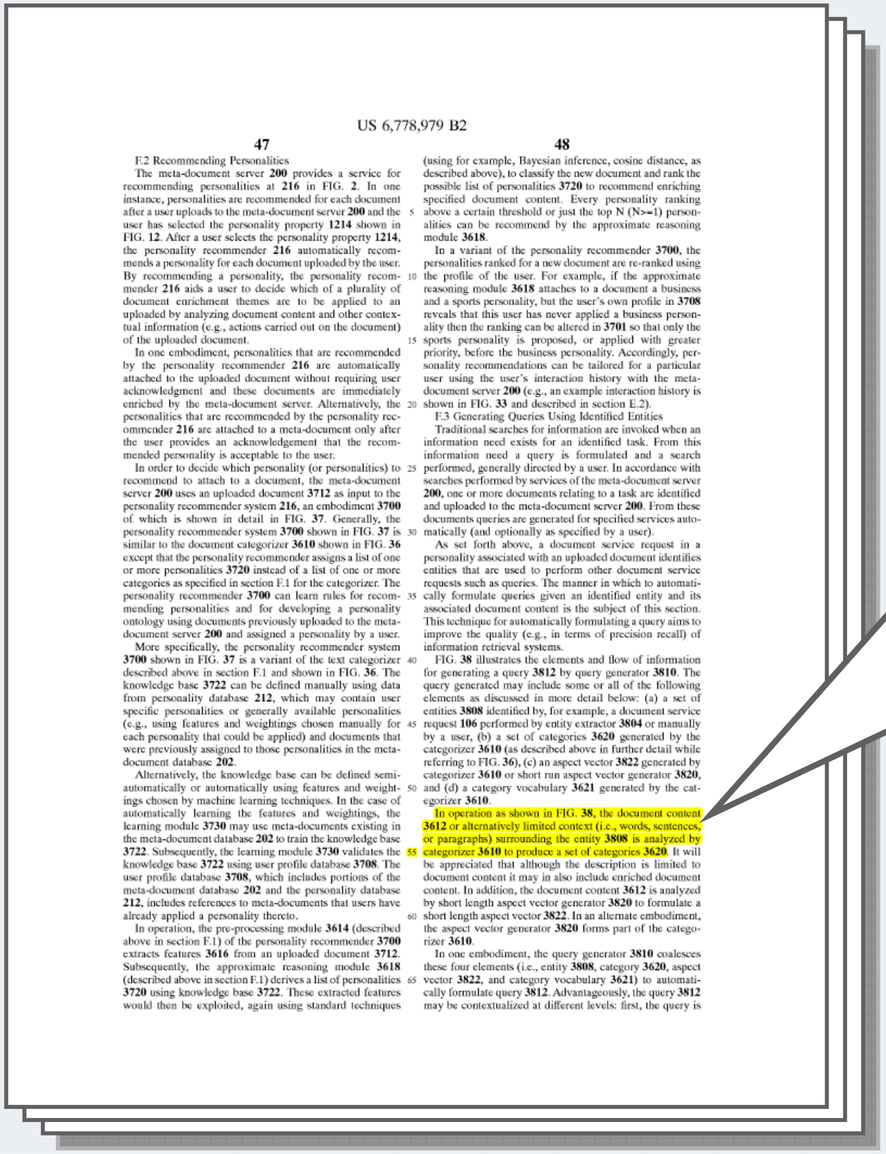
Amendments to the Claims are reflected in the listing of claims which begins on page **2** of this paper.

Remarks begin on page **7** of this paper.

“Applicant’s claims recite automatically generating a query from selected document content, from which both a set of entities and a classification label are automatically identified and assigned, respectively.”

9/8/2003 Applicant’s Amendment at 10; emphasis in original.

"selected document content"



US 6,778,979 B2

47

F.2 Recommending Personalities

The meta-document server **200** provides a service for recommending personalities at **216** in FIG. **2**. In one instance, personalities are recommended for each document after a user uploads to the meta-document server **200** and the user has selected the personality property **1214** shown in FIG. **12**. After a user selects the personality property **1214**, the personality recommender **216** automatically recommends a personality for each document uploaded by the user. By recommending a personality, the personality recommender **216** aids a user to decide which of a plurality of document enrichment themes are to be applied to an uploaded by analyzing document content and other contextual information (e.g., actions carried out on the document) of the uploaded document.

In one embodiment, personalities that are recommended by the personality recommender **216** are automatically attached to the uploaded document without requiring user acknowledgment and these documents are immediately enriched by the meta-document server. Alternatively, the personalities that are recommended by the personality recommender **216** are attached to a meta-document only after the user provides an acknowledgement that the recommended personality is acceptable to the user.

In order to decide which personality (or personalities) to recommend to attach to a document, the meta-document server **200** uses an uploaded document **3712** as input to the personality recommender system **216**, an embodiment **3700** of which is shown in detail in FIG. **37**. Generally, the personality recommender system **3700** shown in FIG. **37** is similar to the document categorizer **3610** shown in FIG. **36** except that the personality recommender assigns a list of one or more personalities **3720** instead of a list of one or more categories as specified in section F.1 for the categorizer. The personality recommender **3700** can learn rules for recommending personalities and for developing a personality ontology using documents previously uploaded to the meta-document server **200** and assigned a personality by a user.

More specifically, the personality recommender system **3700** shown in FIG. **37** is a variant of the text categorizer described above in section F.1 and shown in FIG. **36**. The knowledge base **3722** can be defined manually using data from personality database **212**, which may contain user specific personalities or generally available personalities (e.g., using features and weightings chosen manually for each personality that could be applied) and documents that were previously assigned to those personalities in the meta-document database **202**.

Alternatively, the knowledge base can be defined semi-automatically or automatically using features and weightings chosen by machine learning techniques. In the case of automatically learning the features and weightings, the learning module **3730** may use meta-documents existing in the meta-document database **202** to train the knowledge base **3722**. Subsequently, the learning module **3730** validates the knowledge base **3722** using user profile database **3708**. The user profile database **3708**, which includes portions of the meta-document database **202** and the personality database **212**, includes references to meta-documents that users have already applied a personality thereto.

In operation, the pre-processing module **3614** (described above in section F.1) of the personality recommender **3700** extracts features **3616** from an uploaded document **3712**. Subsequently, the approximate reasoning module **3618** (described above in section F.1) derives a list of personalities **3720** using knowledge base **3722**. These extracted features would then be exploited, again using standard techniques

48

(using for example, Bayesian inference, cosine distance, as described above), to classify the new document and rank the possible list of personalities **3720** to recommend enriching specified document content. Every personality ranking above a certain threshold or just the top N (N>=1) personalities can be recommend by the approximate reasoning module **3618**.

In a variant of the personality recommender **3700**, the personalities ranked for a new document are re-ranked using the profile of the user. For example, if the approximate reasoning module **3618** attaches to a document a business and a sports personality, but the user's own profile in **3708** reveals that this user has never applied a business personality then the ranking can be altered in **3701** so that only the sports personality is proposed, or applied with greater priority, before the business personality. Accordingly, personality recommendations can be tailored for a particular user using the user's interaction history with the meta-document server **200** (e.g., an example interaction history is shown in FIG. **33** and described in section E.2).

F.3 Generating Queries Using Identified Entities

Traditional searches for information are invoked when an information need exists for an identified task. From this information need a query is formulated and a search performed, generally directed by a user. In accordance with searches performed by services of the meta-document server **200**, one or more documents relating to a task are identified and uploaded to the meta-document server **200**. From these documents queries are generated for specified services automatically (and optionally as specified by a user).

As set forth above, a document service request in a personality associated with an uploaded document identifies entities that are used to perform other document service requests such as queries. The manner in which to automatically formulate queries given an identified entity and its associated document content is the subject of this section. This technique for automatically formulating a query aims to improve the quality (e.g., in terms of precision recall) of information retrieval systems.

FIG. **38** illustrates the elements and flow of information for generating a query **3812** by query generator **3810**. The query generated may include some or all of the following elements as discussed in more detail below: (a) a set of entities **3808** identified by, for example, a document service request **106** performed by entity extractor **3804** or manually by a user, (b) a set of categories **3620** generated by the categorizer **3610** (as described above in further detail while referring to FIG. **36**), (c) an aspect vector **3822** generated by categorizer **3610** or short run aspect vector generator **3820**, and (d) a category vocabulary **3621** generated by the categorizer **3610**.

In operation as shown in FIG. **38**, the document content **3612** or alternatively limited context (i.e., words, sentences, or paragraphs) surrounding the entity **3808** is analyzed by categorizer **3610** to produce a set of categories **3620**. It will be appreciated that although the description is limited to document content it may in also include enriched document content. In addition, the document content **3612** is analyzed by short length aspect vector generator **3820** to formulate a short length aspect vector **3822**. In an alternate embodiment, the aspect vector generator **3820** forms part of the categorizer **3610**.

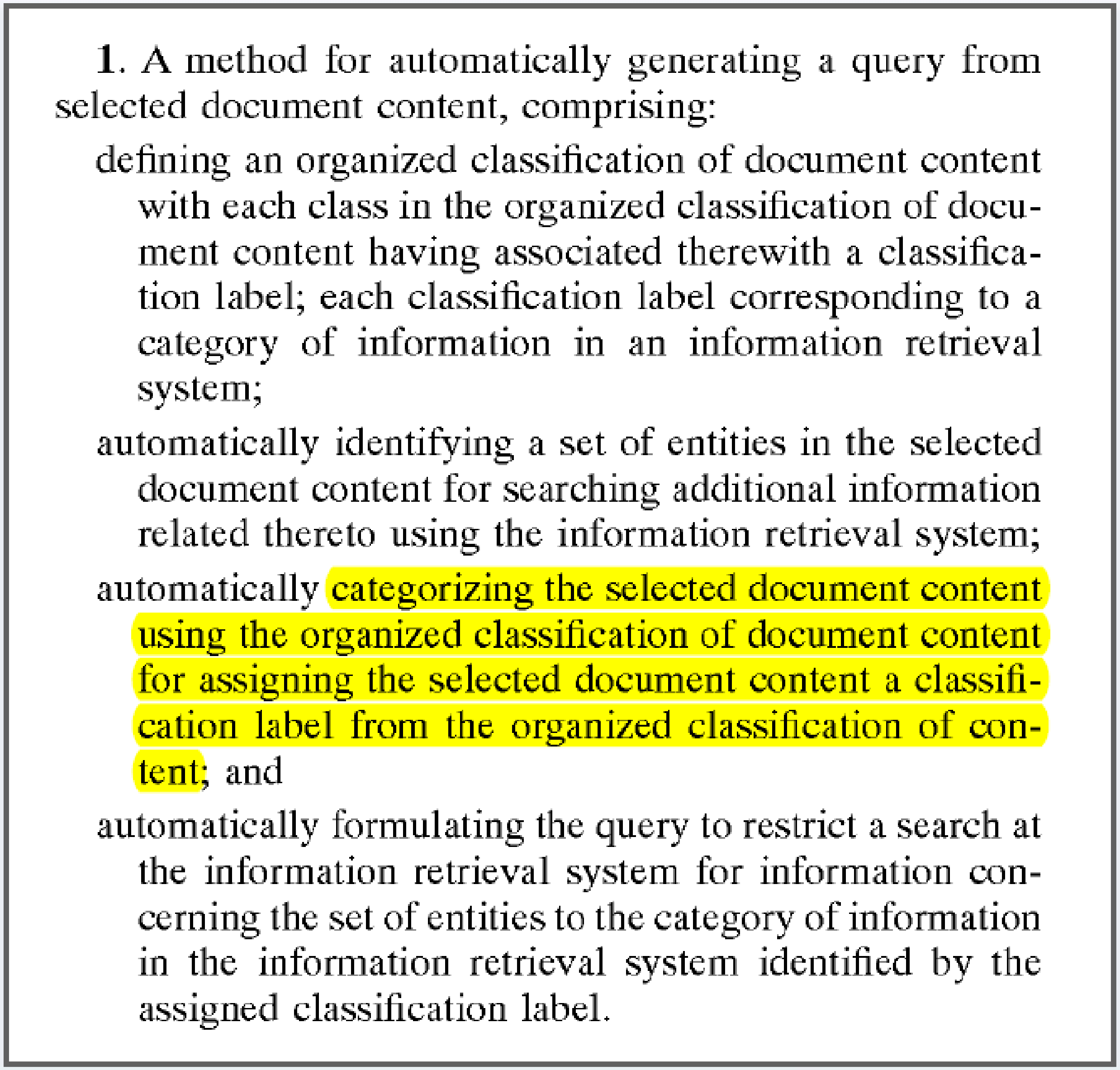
In one embodiment, the query generator **3810** coalesces these four elements (i.e., entity **3808**, category **3620**, aspect vector **3822**, and category vocabulary **3621**) to automatically formulate query **3812**. Advantageously, the query **3812** may be contextualized at different levels: first, the query is

"all or part of the content of a document in electronic form"

"In operation as shown in FIG. 38, the document content 3612 or alternatively limited context (i.e., words, sentences, or paragraphs) surrounding the entity 3808 is analyzed by categorizer 3610 to produce a set of categories 3620."

979/48:52-55; emphasis added

“categorizing the selected document content . . .”



1. A method for automatically generating a query from selected document content, comprising:

defining an organized classification of document content with each class in the organized classification of document content having associated therewith a classification label; each classification label corresponding to a category of information in an information retrieval system;

automatically identifying a set of entities in the selected document content for searching additional information related thereto using the information retrieval system;

automatically categorizing the selected document content using the organized classification of document content for assigning the selected document content a classification label from the organized classification of content; and

automatically formulating the query to restrict a search at the information retrieval system for information concerning the set of entities to the category of information in the information retrieval system identified by the assigned classification label.

"categorizing the selected document content . . ."

Parties' Constructions

Xerox	Defendants
"determining the subject matter of the selected document content using one or more of the categories defining the organized classification of document content and assigning the corresponding classification label(s) to the selected document content."	"using the organized classification of document content to categorize the selected document content and to assign to the selected document content a single classification label."

“categorizing the selected document content . . .”

The “crux of the . . . dispute” according to Defendants

> “The crux of the parties’ dispute is whether, as Defendants contend, a single classification label is assigned in the ‘categorizing’ step and used to identify ‘the’ single category used to restrict a search in the ‘formulating’ step, or whether, as Xerox contends, more than one classification label may be assigned in the ‘categorizing’ step and employed in the ‘formulating’ step.”

Defendants’ Opening Br. at 11.

“categorizing the selected document content . . .”

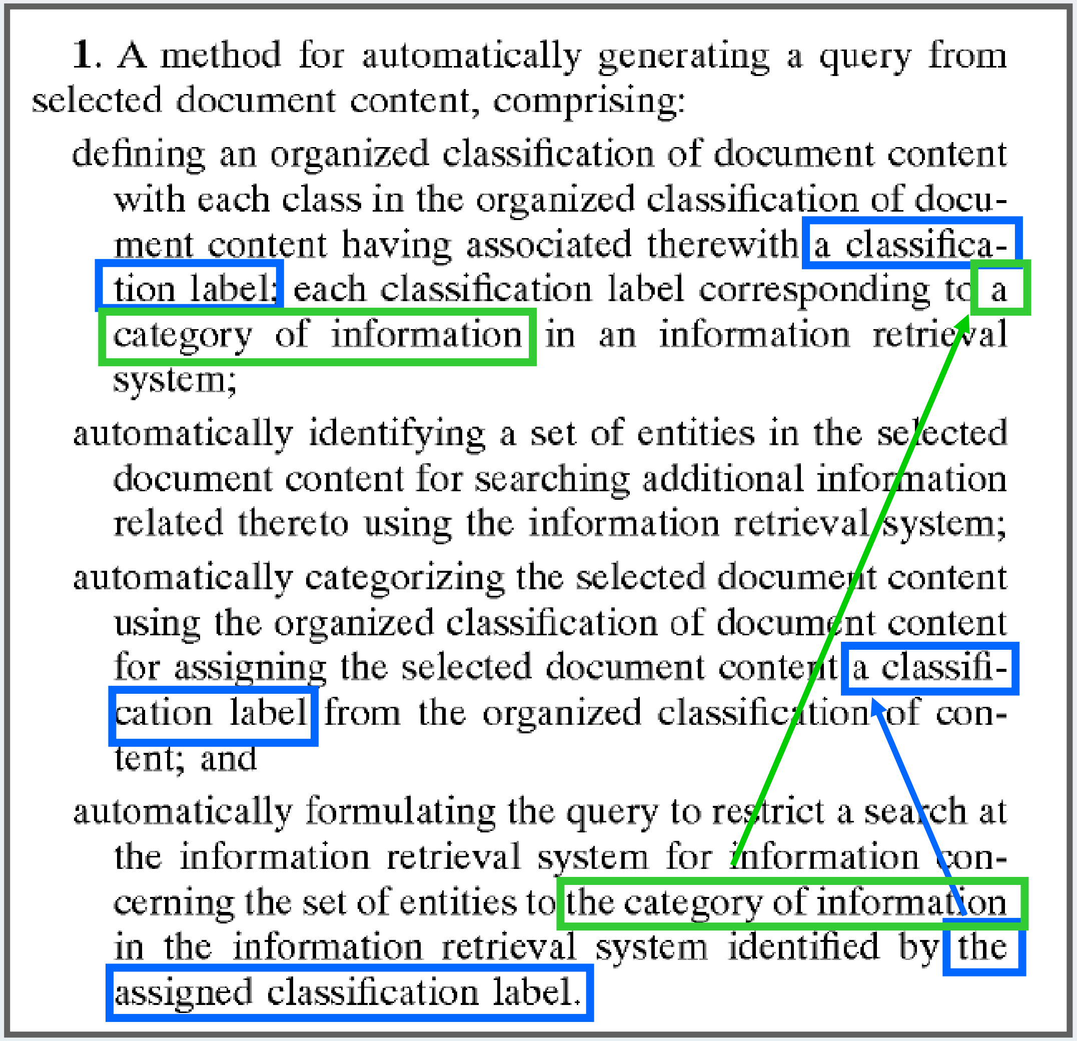
Law on “a” meaning “one or more”

> “That ‘a’ or ‘an’ can mean ‘one or more’ is best described as a rule, rather than merely as a presumption or even a convention. The exceptions to this rule are extremely limited: a patentee must evince a clear intent to limit ‘a’ or ‘an’ to ‘one’.”

Baldwin Graphic Sys., Inc. v. Siebert, Inc., 512 F.3d 1338, 1342-43 (Fed. Cir. 2008) (internal quotations omitted).

“categorizing the selected document content . . .”

Use of one or more labels + categories



1. A method for automatically generating a query from selected document content, comprising:

defining an organized classification of document content with each class in the organized classification of document content having associated therewith a classification label; each classification label corresponding to a category of information in an information retrieval system;

automatically identifying a set of entities in the selected document content for searching additional information related thereto using the information retrieval system;

automatically categorizing the selected document content using the organized classification of document content for assigning the selected document content a classification label from the organized classification of content; and

automatically formulating the query to restrict a search at the information retrieval system for information concerning the set of entities to the category of information in the information retrieval system identified by the assigned classification label.

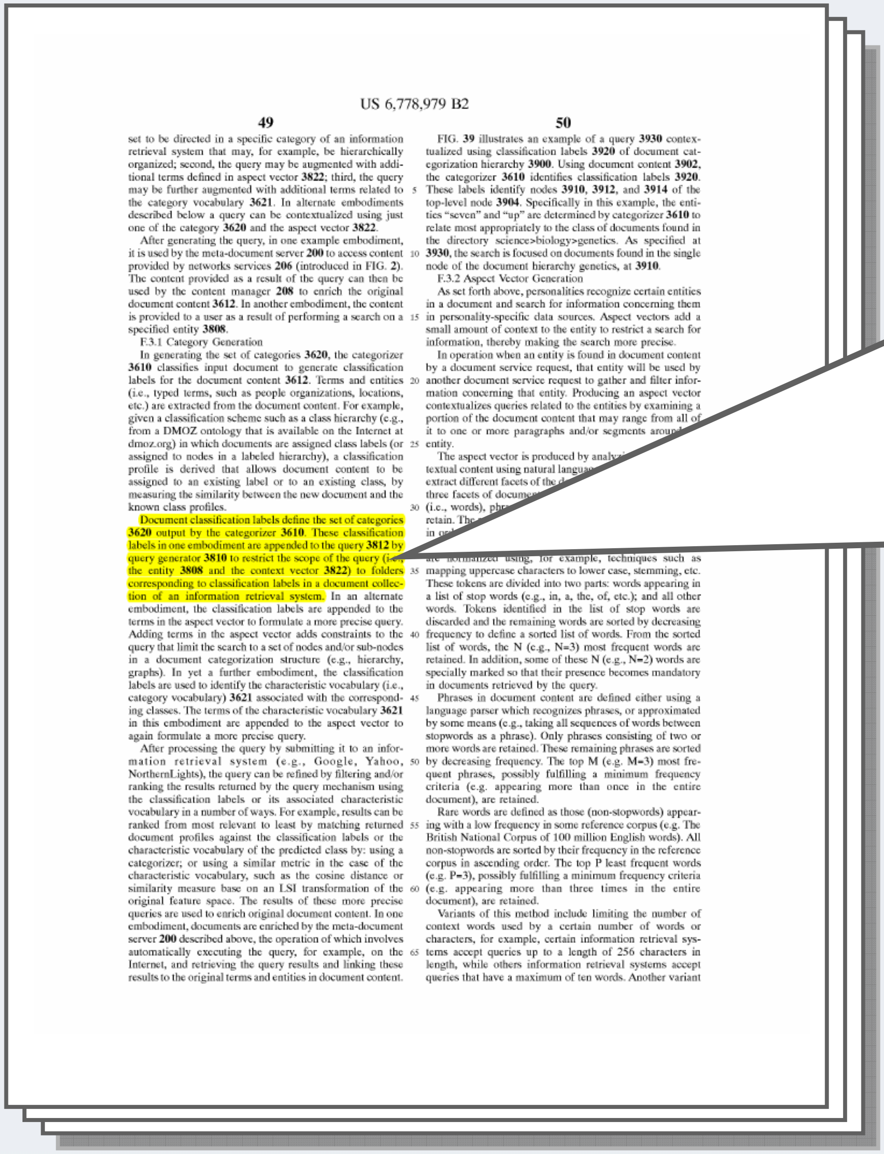
(1) “a” means “one or more”

(2) one or more labels correspond to one or more categories of information in IRS

(3) “the” refers back to one or more labels / categories

(4) search is restricted to one or more categories of information in IRS (identified by one or more labels)

“categorizing the selected document content . . .”



US 6,778,979 B2

49

set to be directed in a specific category of an information retrieval system that may, for example, be hierarchically organized; second, the query may be augmented with additional terms defined in aspect vector **3822**; third, the query may be further augmented with additional terms related to the category vocabulary **3621**. In alternate embodiments described below a query can be contextualized using just one of the category **3620** and the aspect vector **3822**.

After generating the query, in one example embodiment, it is used by the meta-document server **200** to access content provided by networks services **206** (introduced in FIG. **2**). The content provided as a result of the query can then be used by the content manager **208** to enrich the original document content **3612**. In another embodiment, the content is provided to a user as a result of performing a search on a specified entity **3808**.

F.3.1 Category Generation

In generating the set of categories **3620**, the categorizer **3610** classifies input document to generate classification labels for the document content **3612**. Terms and entities (i.e., typed terms, such as people organizations, locations, etc.) are extracted from the document content. For example, given a classification scheme such as a class hierarchy (e.g., from a DMOZ ontology that is available on the Internet at dmoz.org) in which documents are assigned class labels (or assigned to nodes in a labeled hierarchy), a classification profile is derived that allows document content to be assigned to an existing label or to an existing class, by measuring the similarity between the new document and the known class profiles.

Document classification labels define the set of categories **3620** output by the categorizer **3610**. These classification labels in one embodiment are appended to the query **3812** by query generator **3810** to restrict the scope of the query (i.e., the entity **3808** and the context vector **3822**) to folders corresponding to classification labels in a document collection of an information retrieval system. In an alternate embodiment, the classification labels are appended to the terms in the aspect vector to formulate a more precise query. Adding terms in the aspect vector adds constraints to the query that limit the search to a set of nodes and/or sub-nodes in a document categorization structure (e.g., hierarchy, graphs). In yet a further embodiment, the classification labels are used to identify the characteristic vocabulary (i.e., category vocabulary) **3621** associated with the corresponding classes. The terms of the characteristic vocabulary **3621** in this embodiment are appended to the aspect vector to again formulate a more precise query.

After processing the query by submitting it to an information retrieval system (e.g., Google, Yahoo, NorthernLights), the query can be refined by filtering and/or ranking the results returned by the query mechanism using the classification labels or its associated characteristic vocabulary in a number of ways. For example, results can be ranked from most relevant to least by matching returned document profiles against the classification labels or the characteristic vocabulary of the predicted class by: using a categorizer; or using a similar metric in the case of the characteristic vocabulary, such as the cosine distance or similarity measure base on an LSI transformation of the original feature space. The results of these more precise queries are used to enrich original document content. In one embodiment, documents are enriched by the meta-document server **200** described above, the operation of which involves automatically executing the query, for example, on the Internet, and retrieving the query results and linking these results to the original terms and entities in document content.

50

FIG. **39** illustrates an example of a query **3930** contextualized using classification labels **3920** of document categorization hierarchy **3900**. Using document content **3902**, the categorizer **3610** identifies classification labels **3920**. These labels identify nodes **3910**, **3912**, and **3914** of the top-level node **3904**. Specifically in this example, the entities “seven” and “up” are determined by categorizer **3610** to relate most appropriately to the class of documents found in the directory science>biology>genetics. As specified at **3930**, the search is focused on documents found in the single node of the document hierarchy genetics, at **3910**.

F.3.2 Aspect Vector Generation

As set forth above, personalities recognize certain entities in a document and search for information concerning them in personality-specific data sources. Aspect vectors add a small amount of context to the entity to restrict a search for information, thereby making the search more precise.

In operation when an entity is found in document content by a document service request, that entity will be used by another document service request to gather and filter information concerning that entity. Producing an aspect vector contextualizes queries related to the entities by examining a portion of the document content that may range from all of it to one or more paragraphs and/or segments around
entity.

The aspect vector is produced by analy
textual content using natural languag
extract different facets of the d
three facets of documen
(i.e., words), phr
retain. The
in or
are normalized using, for example, techniques such as mapping uppercase characters to lower case, stemming, etc. These tokens are divided into two parts: words appearing in a list of stop words (e.g., in, a, the, of, etc.); and all other words. Tokens identified in the list of stop words are discarded and the remaining words are sorted by decreasing frequency to define a sorted list of words. From the sorted list of words, the N (e.g., N=3) most frequent words are retained. In addition, some of these N (e.g., N=2) words are specially marked so that their presence becomes mandatory in documents retrieved by the query.

Phrases in document content are defined either using a language parser which recognizes phrases, or approximated by some means (e.g., taking all sequences of words between stopwords as a phrase). Only phrases consisting of two or more words are retained. These remaining phrases are sorted by decreasing frequency. The top M (e.g. M=3) most frequent phrases, possibly fulfilling a minimum frequency criteria (e.g. appearing more than once in the entire document), are retained.

Rare words are defined as those (non-stopwords) appearing with a low frequency in some reference corpus (e.g. The British National Corpus of 100 million English words). All non-stopwords are sorted by their frequency in the reference corpus in ascending order. The top P least frequent words (e.g. P=3), possibly fulfilling a minimum frequency criteria (e.g. appearing more than three times in the entire document), are retained.

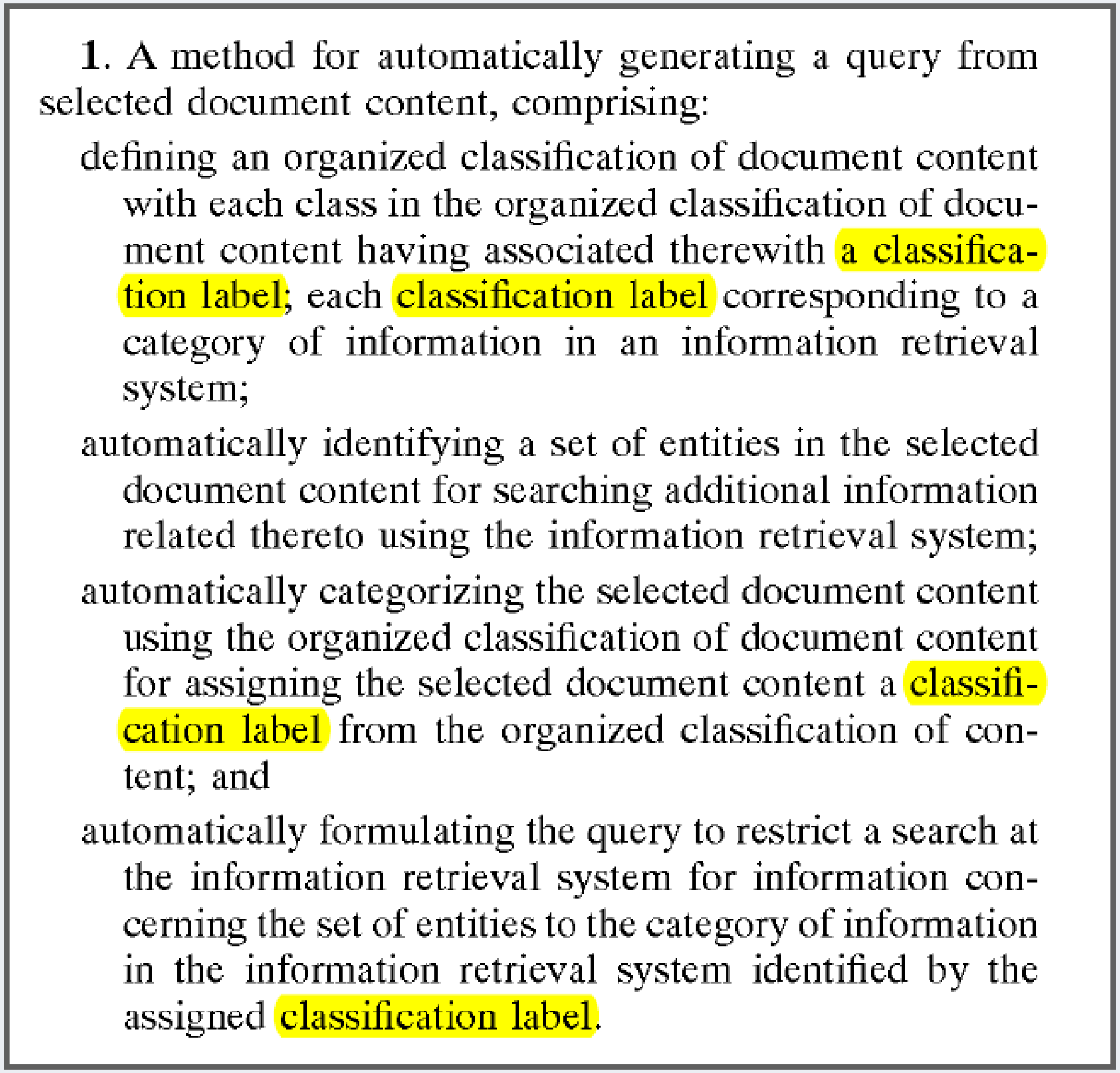
Variants of this method include limiting the number of context words used by a certain number of words or characters, for example, certain information retrieval systems accept queries up to a length of 256 characters in length, while others information retrieval systems accept queries that have a maximum of ten words. Another variant

The ’979 Patent teaches the use of one or more labels.

“Document classification labels define the set of categories 3620 output by the categorizer 3610. These classification labels in one embodiment are appended to the query 3812 by query generator 3810 to restrict the scope of the query (i.e., the entity 3808 and the context vector 3822) to folders corresponding to classification labels in a document collection of an information retrieval system.”

979/49:31-37; emphasis added

“classification label”



1. A method for automatically generating a query from selected document content, comprising:

defining an organized classification of document content with each class in the organized classification of document content having associated therewith a classification label; each classification label corresponding to a category of information in an information retrieval system;

automatically identifying a set of entities in the selected document content for searching additional information related thereto using the information retrieval system;

automatically categorizing the selected document content using the organized classification of document content for assigning the selected document content a classification label from the organized classification of content; and

automatically formulating the query to restrict a search at the information retrieval system for information concerning the set of entities to the category of information in the information retrieval system identified by the assigned classification label.

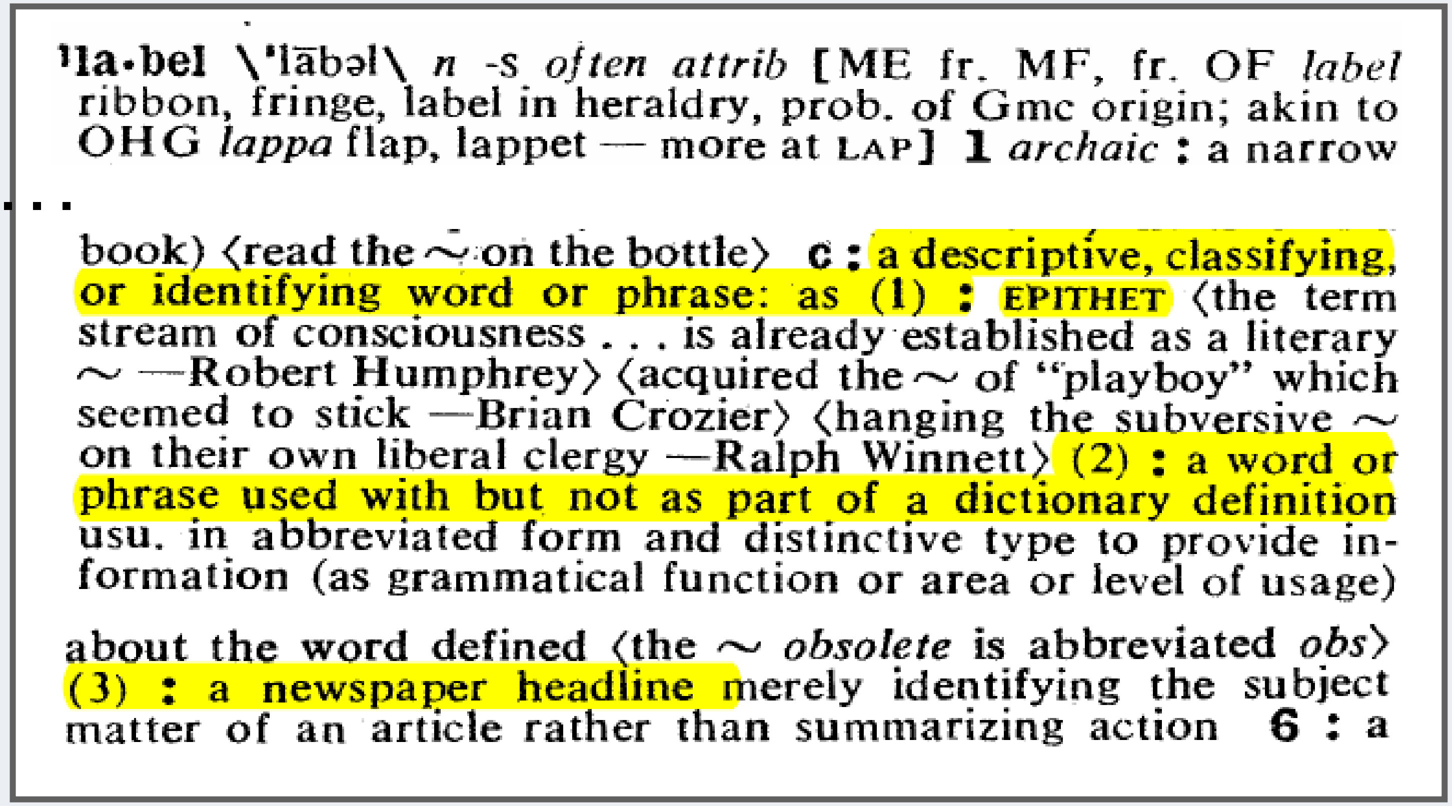
“classification label”

Parties’ Constructions

Xerox	Defendants
“a label in any format that identifies a category in the organized classification of document content.”	“classifying word or phrase”

“classification label”

Example of Defendants’ General Dictionary Definitions



¹la·bel \ˈlābəl\ *n* -s *often attrib* [ME fr. MF, fr. OF *label* ribbon, fringe, label in heraldry, prob. of Gmc origin; akin to OHG *lappa* flap, lappet — more at LAP] **1** *archaic* **:** a narrow

. . .

book) ⟨read the ~ on the bottle⟩ **c :** a descriptive, classifying, or identifying word or phrase: as (1) **:** EPITHET ⟨the term stream of consciousness . . . is already established as a literary ~ —Robert Humphrey⟩ ⟨acquired the ~ of “playboy” which seemed to stick —Brian Crozier⟩ ⟨hanging the subversive ~ on their own liberal clergy —Ralph Winnett⟩ (2) **:** a word or phrase used with but not as part of a dictionary definition usu. in abbreviated form and distinctive type to provide information (as grammatical function or area or level of usage)

about the word defined ⟨the ~ *obsolete* is abbreviated *obs*⟩ (3) **:** a newspaper headline merely identifying the subject matter of an article rather than summarizing action **6 :** a

Webster’s Third New Int’l Dictionary, Unabridged (2002)

“classification label”

Law on Use of General Dictionaries

“By design, general dictionaries collect the definitions of a term as used not only in a particular art field, but in many different settings. . . . For that reason, we have stated that a general-usage dictionary cannot overcome art-specific evidence of the meaning of a claim term.”

Phillips v. AWH Corp., 415 F.3d 1303, 1321-22 (Fed. Cir. 2005).

“classification label”

Examples of Art-Specific Dictionaries

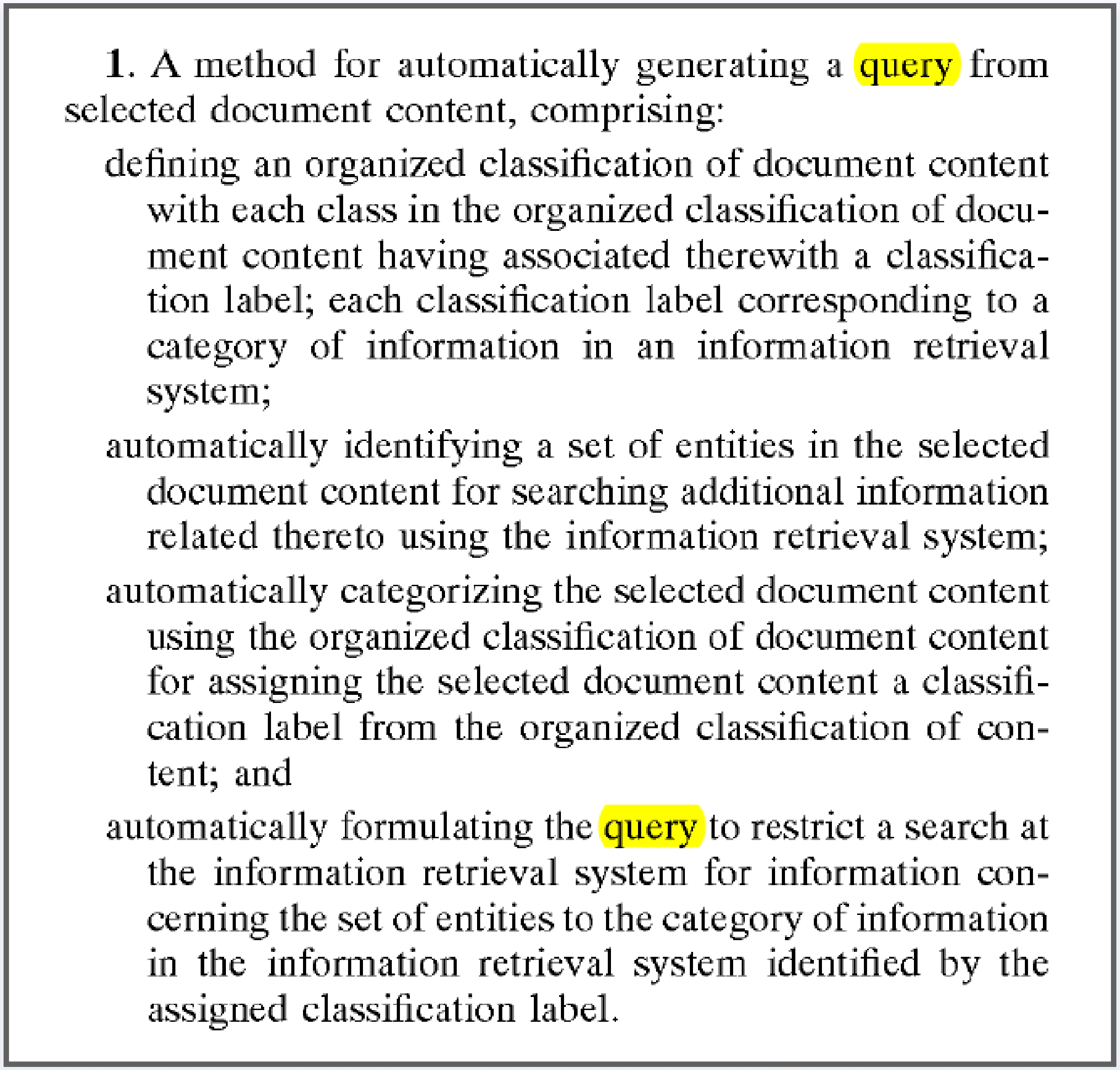
“[a]n identifier within or attached to a set of data elements.”

IBM Dictionary of Computing (1994).

“[a] data item that serves to identify a data record (much in the same way as a key is used) . . . .”

McGraw-Hill Dictionary of Computing & Communications (2003)

“query”



1. A method for automatically generating a query from selected document content, comprising:

defining an organized classification of document content with each class in the organized classification of document content having associated therewith a classification label; each classification label corresponding to a category of information in an information retrieval system;

automatically identifying a set of entities in the selected document content for searching additional information related thereto using the information retrieval system;

automatically categorizing the selected document content using the organized classification of document content for assigning the selected document content a classification label from the organized classification of content; and

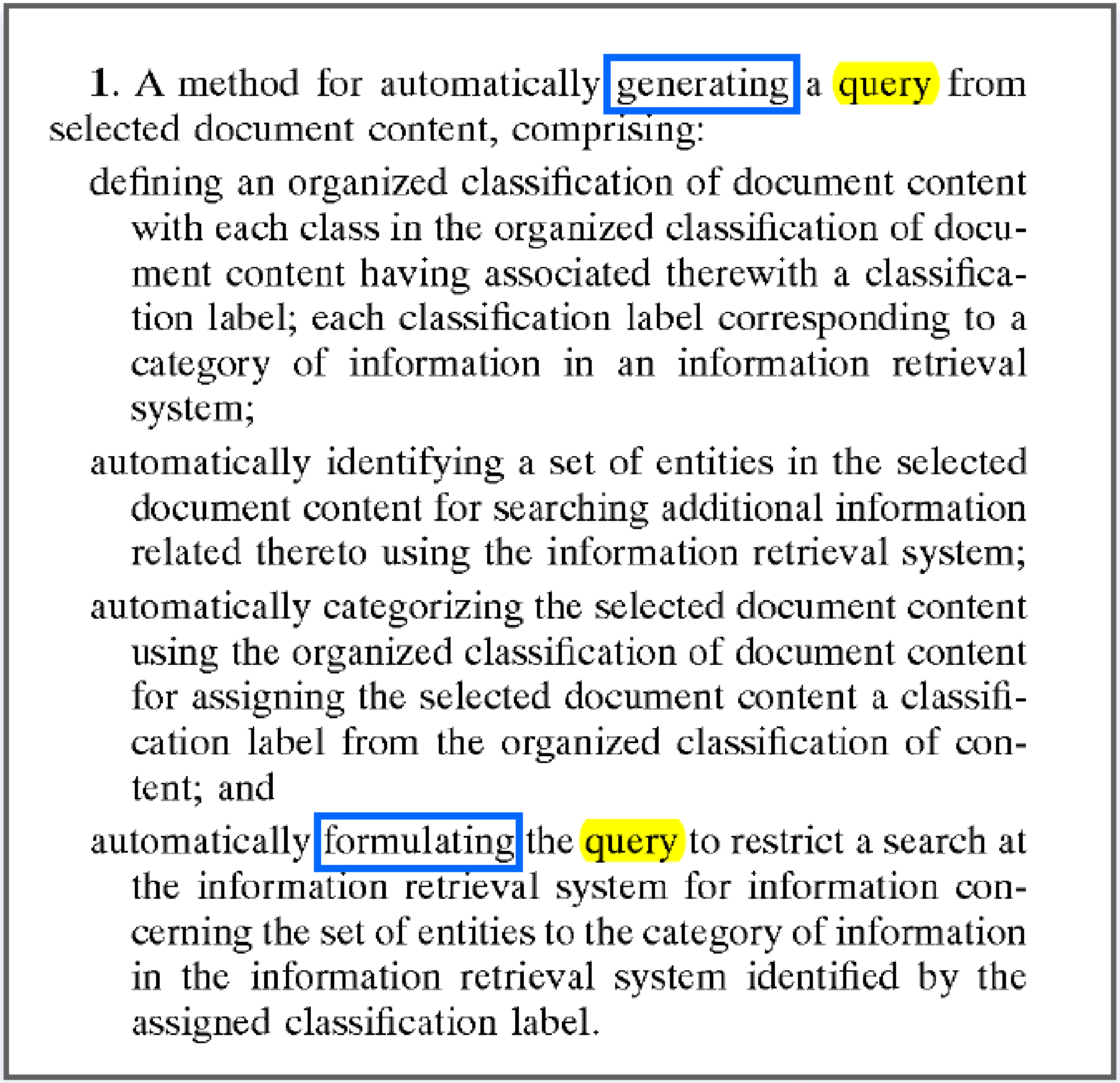
automatically formulating the query to restrict a search at the information retrieval system for information concerning the set of entities to the category of information in the information retrieval system identified by the assigned classification label.

“query”

Parties’ Constructions

Xerox	Defendants
“a set of data specifying search criteria”	“request for search results”

“query”



1. A method for automatically generating a query from selected document content, comprising:

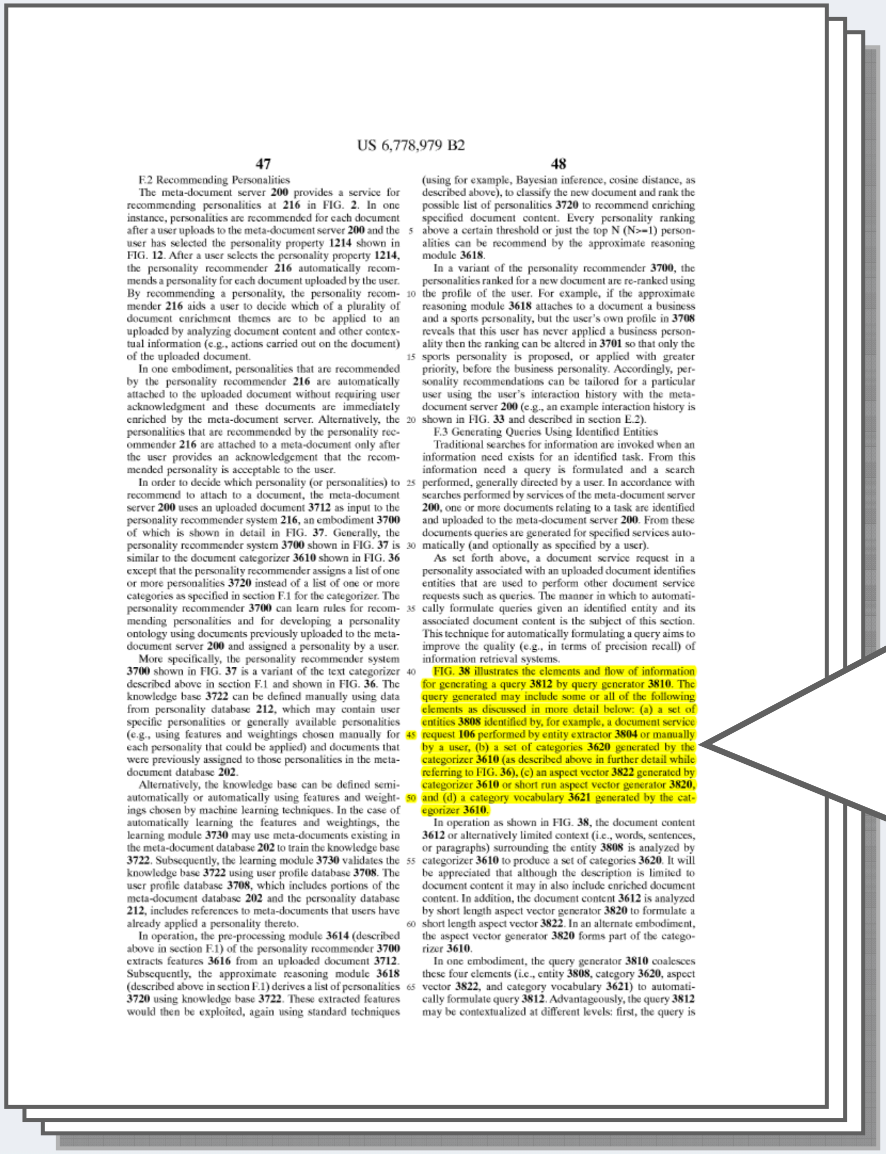
defining an organized classification of document content with each class in the organized classification of document content having associated therewith a classification label; each classification label corresponding to a category of information in an information retrieval system;

automatically identifying a set of entities in the selected document content for searching additional information related thereto using the information retrieval system;

automatically categorizing the selected document content using the organized classification of document content for assigning the selected document content a classification label from the organized classification of content; and

automatically formulating the query to restrict a search at the information retrieval system for information concerning the set of entities to the category of information in the information retrieval system identified by the assigned classification label.

"query"



US 6,778,979 B2

47

F.2 Recommending Personalities

The meta-document server **200** provides a service for recommending personalities at **216** in FIG. **2**. In one instance, personalities are recommended for each document after a user uploads to the meta-document server **200** and the user has selected the personality property **1214** shown in FIG. **12**. After a user selects the personality property **1214**, the personality recommender **216** automatically recommends a personality for each document uploaded by the user. By recommending a personality, the personality recommender **216** aids a user to decide which of a plurality of document enrichment themes are to be applied to an uploaded by analyzing document content and other contextual information (e.g., actions carried out on the document) of the uploaded document.

In one embodiment, personalities that are recommended by the personality recommender **216** are automatically attached to the uploaded document without requiring user acknowledgment and these documents are immediately enriched by the meta-document server. Alternatively, the personalities that are recommended by the personality recommender **216** are attached to a meta-document only after the user provides an acknowledgement that the recommended personality is acceptable to the user.

In order to decide which personality (or personalities) to recommend to attach to a document, the meta-document server **200** uses an uploaded document **3712** as input to the personality recommender system **216**, an embodiment **3700** of which is shown in detail in FIG. **37**. Generally, the personality recommender system **3700** shown in FIG. **37** is similar to the document categorizer **3610** shown in FIG. **36** except that the personality recommender assigns a list of one or more personalities **3720** instead of a list of one or more categories as specified in section F.1 for the categorizer. The personality recommender **3700** can learn rules for recommending personalities and for developing a personality ontology using documents previously uploaded to the meta-document server **200** and assigned a personality by a user.

More specifically, the personality recommender system **3700** shown in FIG. **37** is a variant of the text categorizer described above in section F.1 and shown in FIG. **36**. The knowledge base **3722** can be defined manually using data from personality database **212**, which may contain user specific personalities or generally available personalities (e.g., using features and weightings chosen manually for each personality that could be applied) and documents that were previously assigned to those personalities in the meta-document database **202**.

Alternatively, the knowledge base can be defined semi-automatically or automatically using features and weightings chosen by machine learning techniques. In the case of automatically learning the features and weightings, the learning module **3730** may use meta-documents existing in the meta-document database **202** to train the knowledge base **3722**. Subsequently, the learning module **3730** validates the knowledge base **3722** using user profile database **3708**. The user profile database **3708**, which includes portions of the meta-document database **202** and the personality database **212**, includes references to meta-documents that users have already applied a personality thereto.

In operation, the pre-processing module **3614** (described above in section F.1) of the personality recommender **3700** extracts features **3616** from an uploaded document **3712**. Subsequently, the approximate reasoning module **3618** (described above in section F.1) derives a list of personalities **3720** using knowledge base **3722**. These extracted features would then be exploited, again using standard techniques

48

(using for example, Bayesian inference, cosine distance, as described above), to classify the new document and rank the possible list of personalities **3720** to recommend enriching specified document content. Every personality ranking above a certain threshold or just the top N (N>=1) personalities can be recommend by the approximate reasoning module **3618**.

In a variant of the personality recommender **3700**, the personalities ranked for a new document are re-ranked using the profile of the user. For example, if the approximate reasoning module **3618** attaches to a document a business and a sports personality, but the user's own profile in **3708** reveals that this user has never applied a business personality then the ranking can be altered in **3701** so that only the sports personality is proposed, or applied with greater priority, before the business personality. Accordingly, personality recommendations can be tailored for a particular user using the user's interaction history with the meta-document server **200** (e.g., an example interaction history is shown in FIG. **33** and described in section E.2).

F.3 Generating Queries Using Identified Entities

Traditional searches for information are invoked when an information need exists for an identified task. From this information need a query is formulated and a search performed, generally directed by a user. In accordance with searches performed by services of the meta-document server **200**, one or more documents relating to a task are identified and uploaded to the meta-document server **200**. From these documents queries are generated for specified services automatically (and optionally as specified by a user).

As set forth above, a document service request in a personality associated with an uploaded document identifies entities that are used to perform other document service requests such as queries. The manner in which to automatically formulate queries given an identified entity and its associated document content is the subject of this section. This technique for automatically formulating a query aims to improve the quality (e.g., in terms of precision recall) of information retrieval systems.

FIG. **38** illustrates the elements and flow of information for generating a query **3812** by query generator **3810**. The query generated may include some or all of the following elements as discussed in more detail below: (a) a set of entities **3808** identified by, for example, a document service request **106** performed by entity extractor **3804** or manually by a user, (b) a set of categories **3620** generated by the categorizer **3610** (as described above in further detail while referring to FIG. **36**), (c) an aspect vector **3822** generated by categorizer **3610** or short run aspect vector generator **3820**, and (d) a category vocabulary **3621** generated by the categorizer **3610**.

In operation as shown in FIG. **38**, the document content **3612** or alternatively limited context (i.e., words, sentences, or paragraphs) surrounding the entity **3808** is analyzed by categorizer **3610** to produce a set of categories **3620**. It will be appreciated that although the description is limited to document content it may in also include enriched document content. In addition, the document content **3612** is analyzed by short length aspect vector generator **3820** to formulate a short length aspect vector **3822**. In an alternate embodiment, the aspect vector generator **3820** forms part of the categorizer **3610**.

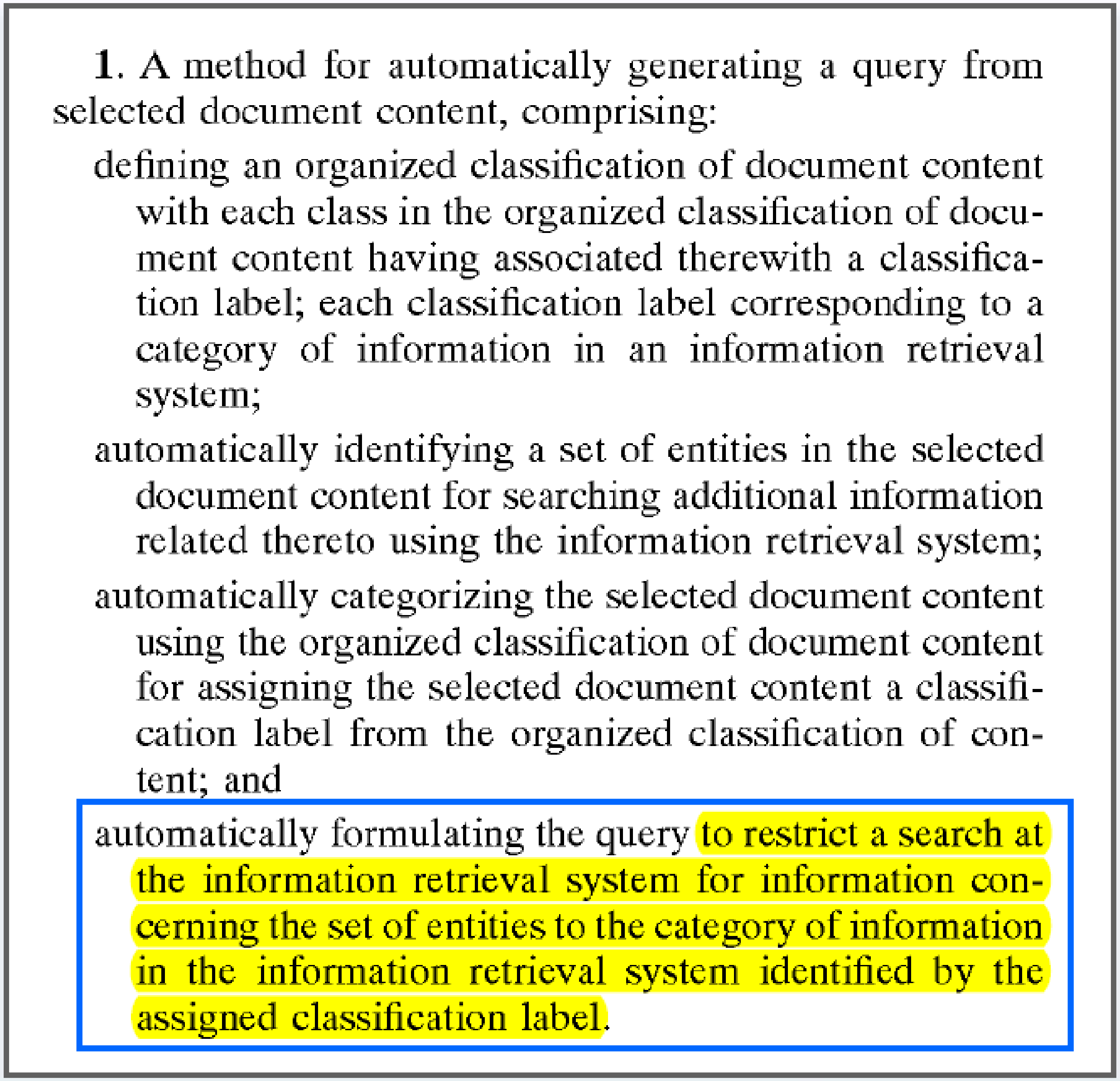
In one embodiment, the query generator **3810** coalesces these four elements (i.e., entity **3808**, category **3620**, aspect vector **3822**, and category vocabulary **3621**) to automatically formulate query **3812**. Advantageously, the query **3812** may be contextualized at different levels: first, the query is

A "query" in the '979 Patent is defined by its contents

"The query generated may include some or all of the following elements…: (a) a set of entities 3808…, (b) a set of categories 3620 generated by the categorizer 3610…, (c) an aspect vector 3822 generated by categorizer 3610 or short run aspect vector generator 3820, and (d) a category vocabulary 3621 generated by the categorizer 3610"

979/48:41-51

“to restrict . . . label”



1. A method for automatically generating a query from selected document content, comprising:

defining an organized classification of document content with each class in the organized classification of document content having associated therewith a classification label; each classification label corresponding to a category of information in an information retrieval system;

automatically identifying a set of entities in the selected document content for searching additional information related thereto using the information retrieval system;

automatically categorizing the selected document content using the organized classification of document content for assigning the selected document content a classification label from the organized classification of content; and

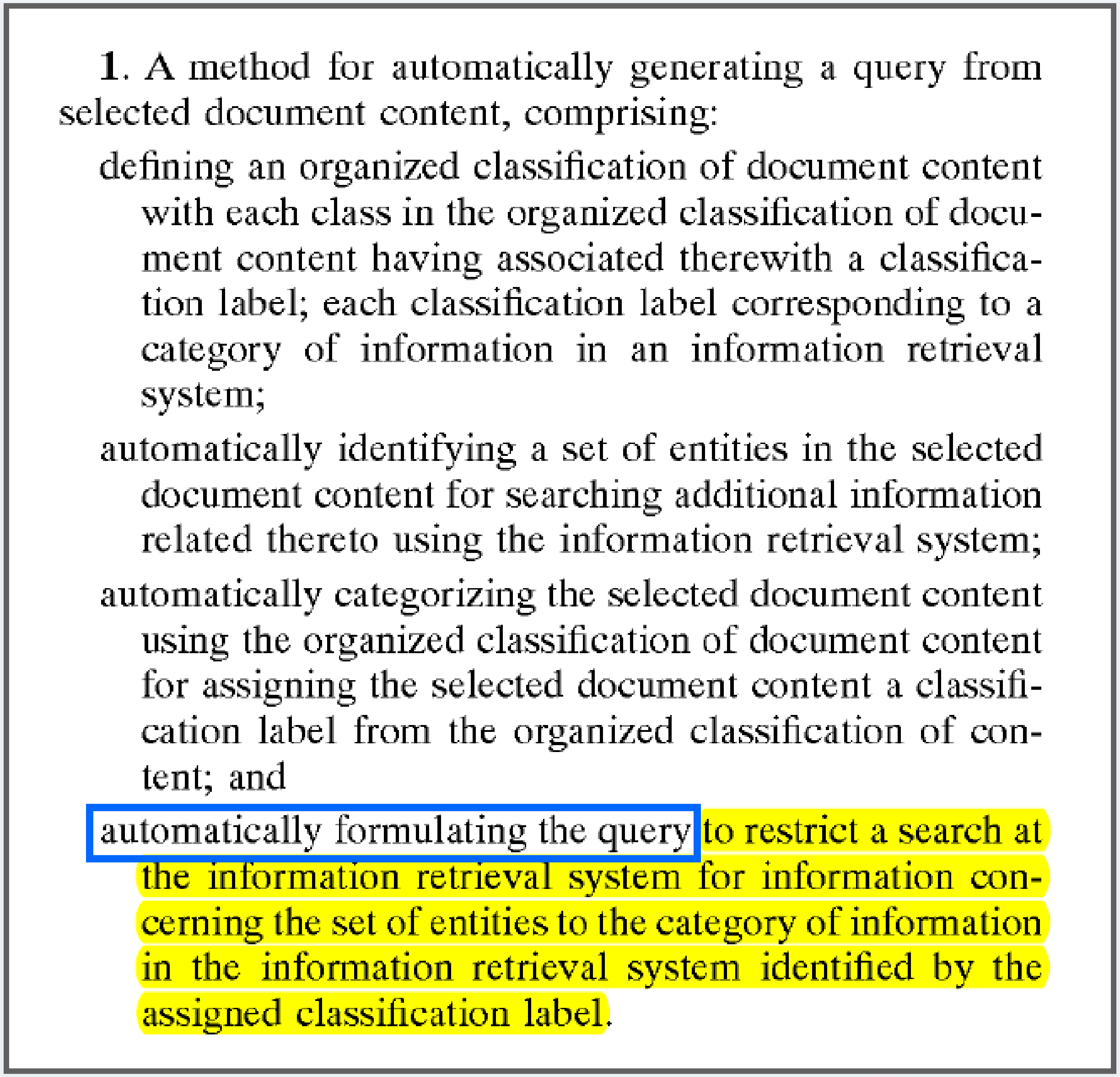
automatically formulating the query to restrict a search at the information retrieval system for information concerning the set of entities to the category of information in the information retrieval system identified by the assigned classification label.

“to restrict . . . label”

Parties’ Constructions

Xerox	Defendants
“the set of data specifying search criteria includes data items corresponding to one or more entities identified in the ‘automatically identifying’ step and one or more classification labels assigned in the ‘automatically categorizing’ step.”	“to confine a search at the information retrieval system to the category of information identified by the assigned classification label, where the search seeks information concerning the set of entities.”

“to restrict . . . label”



1. A method for automatically generating a query from selected document content, comprising:

defining an organized classification of document content with each class in the organized classification of document content having associated therewith a classification label; each classification label corresponding to a category of information in an information retrieval system;

automatically identifying a set of entities in the selected document content for searching additional information related thereto using the information retrieval system;

automatically categorizing the selected document content using the organized classification of document content for assigning the selected document content a classification label from the organized classification of content; and

automatically formulating the query to restrict a search at the information retrieval system for information concerning the set of entities to the category of information in the information retrieval system identified by the assigned classification label.

“to restrict . . . label”

The specification portion Defendants cite confirms that Xerox’s construction is correct, despite Defendants’ misleading underlining.

> “The specification explains that ‘the search is focused on documents found in the single node of the document hierarchy genetics, at 3910’.”

Defs.’ Br. at 8 (quoting 979/50:10-11); emphasis in original.

Order of Steps (Claims 1, 18)

Parties’ Constructions (Claims 1, 18)

Xerox	Defendants
Step (a) must be performed before steps (c) and (d).	Step (a) must be performed before steps (c) and (d).
Step (b) must be performed before the completion of step (d).	Step (b) must be performed before step (d).
Step (c) must be performed before the completion of step (d).	Step (c) must be performed before step (d).

Order of Steps (Claims 1, 18)

Law on order of steps

> “[A]s a general rule the claim is not limited to performance of the steps in the order recited, unless the claim explicitly or implicitly *requires* a specific order.”

Baldwin Graphic Sys., Inc. v. Siebert, Inc., 512 F.3d 1338, 1345 (Fed. Cir. 2008) (emphasis added).

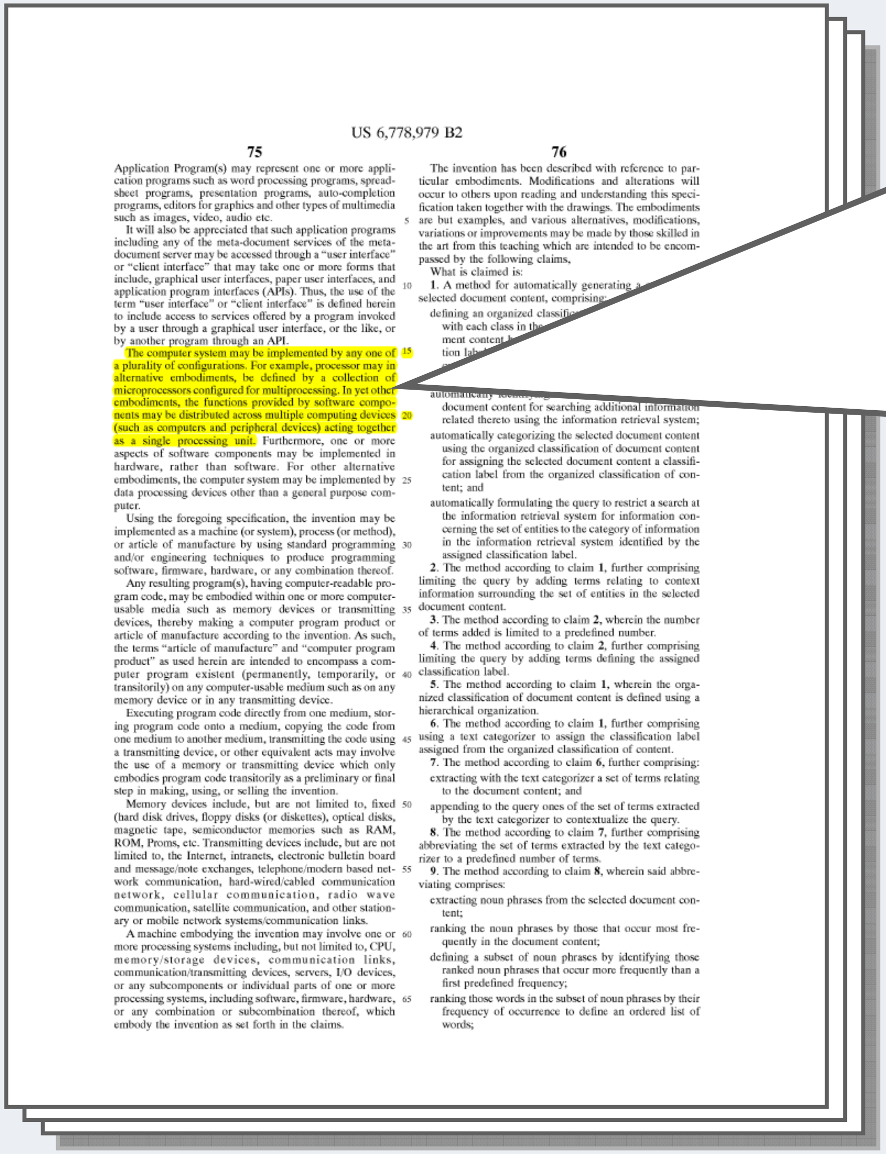
Order of Steps (Claims 1, 18)

Law on order of steps

> “First, we look to the claim language to determine if, as a matter of logic or grammar, they must be performed in the order written. . . . If not, we next look to the rest of the specification to determine whether it directly or implicitly requires such a narrow construction. If not, the sequence in which such steps are written is not a requirement.”

Altiris, Inc. v. Symantec Corp., 318 F.3d 1363, 1369-70 (Fed. Cir. 2003) (internal citations omitted).

Order of Steps (Claims 1, 18)



US 6,778,979 B2

75

Application Program(s) may represent one or more application programs such as word processing programs, spreadsheet programs, presentation programs, auto-completion programs, editors for graphics and other types of multimedia such as images, video, audio etc.

It will also be appreciated that such application programs including any of the meta-document services of the meta-document server may be accessed through a "user interface" or "client interface" that may take one or more forms that include, graphical user interfaces, paper user interfaces, and application program interfaces (APIs). Thus, the use of the term "user interface" or "client interface" is defined herein to include access to services offered by a program invoked by a user through a graphical user interface, or the like, or by another program through an API.

The computer system may be implemented by any one of a plurality of configurations. For example, processor may in alternative embodiments, be defined by a collection of microprocessors configured for multiprocessing. In yet other embodiments, the functions provided by software components may be distributed across multiple computing devices (such as computers and peripheral devices) acting together as a single processing unit. Furthermore, one or more aspects of software components may be implemented in hardware, rather than software. For other alternative embodiments, the computer system may be implemented by data processing devices other than a general purpose computer.

Using the foregoing specification, the invention may be implemented as a machine (or system), process (or method), or article of manufacture by using standard programming and/or engineering techniques to produce programming software, firmware, hardware, or any combination thereof.

Any resulting program(s), having computer-readable program code, may be embodied within one or more computer-usable media such as memory devices or transmitting devices, thereby making a computer program product or article of manufacture according to the invention. As such, the terms "article of manufacture" and "computer program product" as used herein are intended to encompass a computer program existent (permanently, temporarily, or transitorily) on any computer-usable medium such as on any memory device or in any transmitting device.

Executing program code directly from one medium, storing program code onto a medium, copying the code from one medium to another medium, transmitting the code using a transmitting device, or other equivalent acts may involve the use of a memory or transmitting device which only embodies program code transitorily as a preliminary or final step in making, using, or selling the invention.

Memory devices include, but are not limited to, fixed (hard disk drives, floppy disks (or diskettes), optical disks, magnetic tape, semiconductor memories such as RAM, ROM, Proms, etc. Transmitting devices include, but are not limited to, the Internet, intranets, electronic bulletin board and message/note exchanges, telephone/modern based network communication, hard-wired/cabled communication network, cellular communication, radio wave communication, satellite communication, and other stationary or mobile network systems/communication links.

A machine embodying the invention may involve one or more processing systems including, but not limited to, CPU, memory/storage devices, communication links, communication/transmitting devices, servers, I/O devices, or any subcomponents or individual parts of one or more processing systems, including software, firmware, hardware, or any combination or subcombination thereof, which embody the invention as set forth in the claims.

76

The invention has been described with reference to particular embodiments. Modifications and alterations will occur to others upon reading and understanding this specification taken together with the drawings. The embodiments are but examples, and various alternatives, modifications, variations or improvements may be made by those skilled in the art from this teaching which are intended to be encompassed by the following claims,

What is claimed is:

1. A method for automatically generating a

selected document content, comprising:

defining an organized classific

with each class in th

ment content h

tion lab

automatically

document content for searching additional information related thereto using the information retrieval system;

automatically categorizing the selected document content using the organized classification of document content for assigning the selected document content a classification label from the organized classification of content; and

automatically formulating the query to restrict a search at the information retrieval system for information concerning the set of entities to the category of information in the information retrieval system identified by the assigned classification label.

2. The method according to claim **1**, further comprising limiting the query by adding terms relating to context information surrounding the set of entities in the selected document content.

3. The method according to claim **2**, wherein the number of terms added is limited to a predefined number.

4. The method according to claim **2**, further comprising limiting the query by adding terms defining the assigned classification label.

5. The method according to claim **1**, wherein the organized classification of document content is defined using a hierarchical organization.

6. The method according to claim **1**, further comprising using a text categorizer to assign the classification label assigned from the organized classification of content.

7. The method according to claim **6**, further comprising:

extracting with the text categorizer a set of terms relating to the document content; and

appending to the query ones of the set of terms extracted by the text categorizer to contextualize the query.

8. The method according to claim **7**, further comprising abbreviating the set of terms extracted by the text categorizer to a predefined number of terms.

9. The method according to claim **8**, wherein said abbreviating comprises:

extracting noun phrases from the selected document content;

ranking the noun phrases by those that occur most frequently in the document content;

defining a subset of noun phrases by identifying those ranked noun phrases that occur more frequently than a first predefined frequency;

ranking those words in the subset of noun phrases by their frequency of occurrence to define an ordered list of words;

"The computer system may be implemented by any one of a plurality of configurations. For example, processor may in alternative embodiments, be defined by a collection of microprocessors configured for multiprocessing. In yet other embodiments, the functions provided by software components may be distributed across multiple computing devices (such as computers and peripheral devices) acting together as a single processing unit."

979/75:15-22

Order of Steps (Claims 1 & 2; 18 & 19)

Parties’ Constructions (Claims 1 & 2)

Xerox	Defendants
The step of Claim 2 must be performed during or after the completion of step (d) of Claim 1.	The steps of claim 1 must be performed before the step of 2.

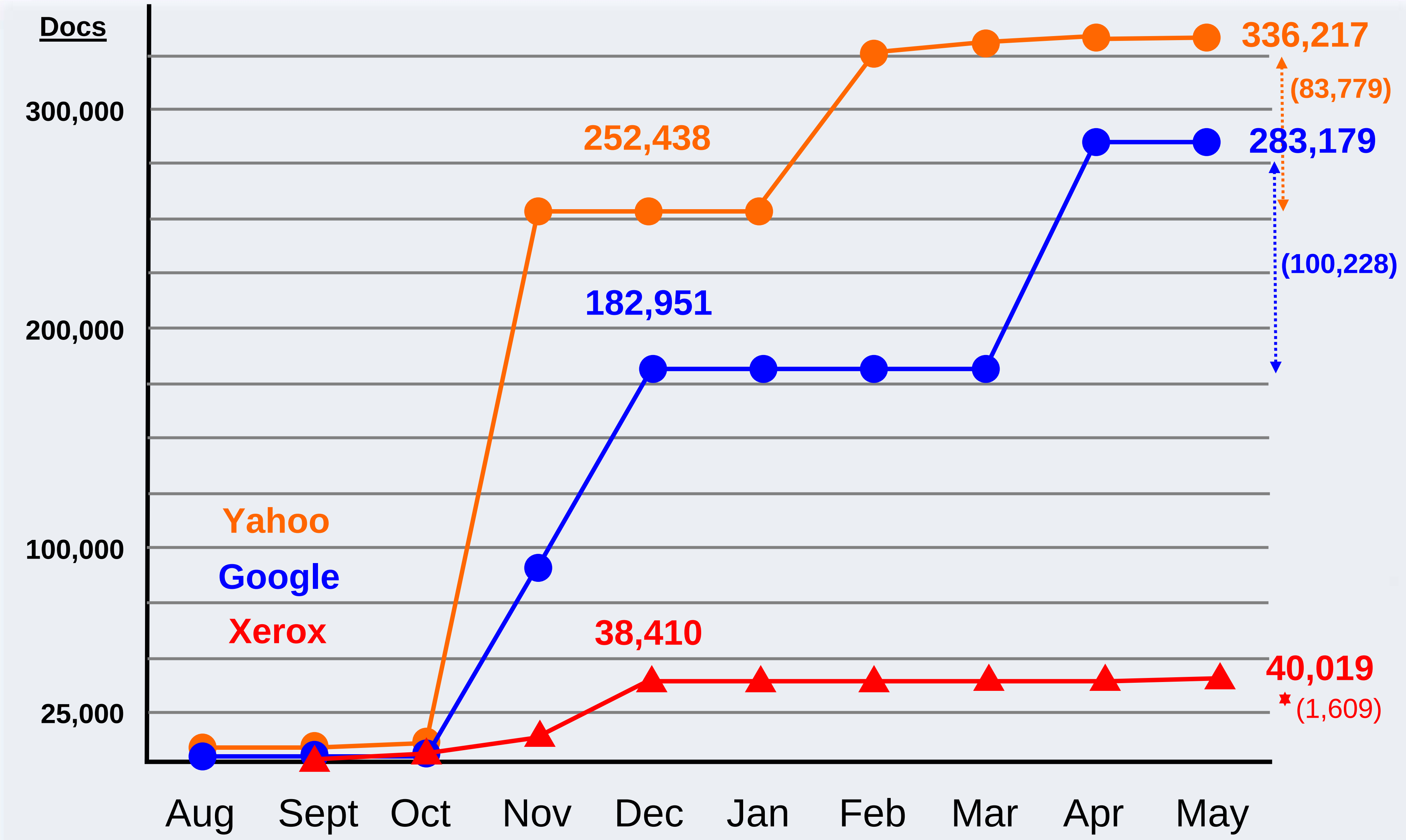
Order of Steps (Claims 1 & 2; 18 & 19)

“Entities” (Claim 1) and “terms” (Claim 2) can be present in the formulated query before “categories” (Claim 1).

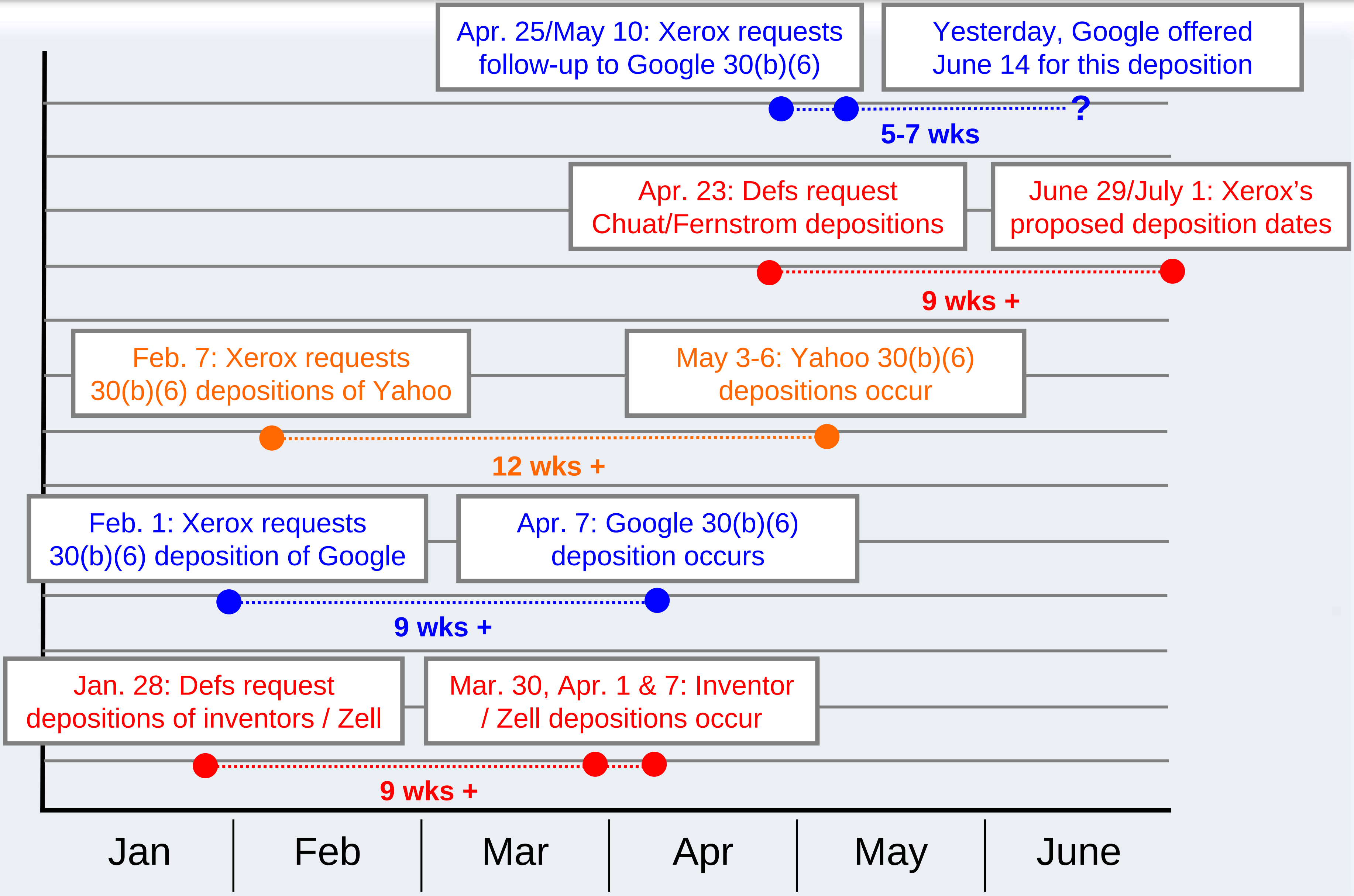
> “Document classification labels define the set of categories 3620 output by the categorizer 3610. These classification labels in one embodiment are appended to the query 3812 by query generator 3810 to restrict the scope of the query (i.e., the entity 3808 and the context vector 3822)”

979/49:31-35; emphasis added.

The Parties' Production History



The Parties’ Deposition History



Apr. 25/May 10: Xerox requests follow-up to Google 30(b)(6)

Yesterday, Google offered June 14 for this deposition

Apr. 23: Defs request Chuat/Fernstrom depositions

June 29/July 1: Xerox’s proposed deposition dates

Feb. 7: Xerox requests 30(b)(6) depositions of Yahoo

May 3-6: Yahoo 30(b)(6) depositions occur

Feb. 1: Xerox requests 30(b)(6) deposition of Google

Apr. 7: Google 30(b)(6) deposition occurs

Jan. 28: Defs request depositions of inventors / Zell

Mar. 30, Apr. 1 & 7: Inventor / Zell depositions occur