

EXHIBIT A

Introduction / The Problem

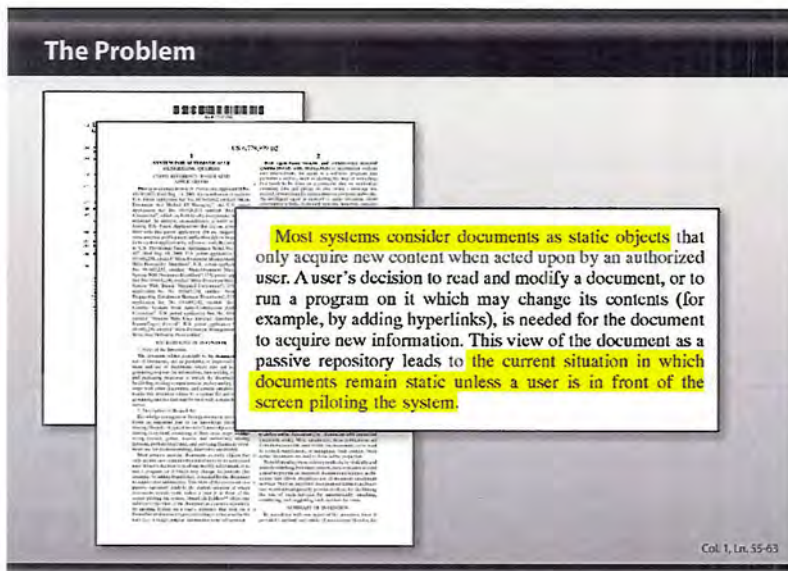


May it please the Court:

Defendants Google, Yahoo!, and Right Media present this technology tutorial in the patent infringement action brought by Plaintiff Xerox.

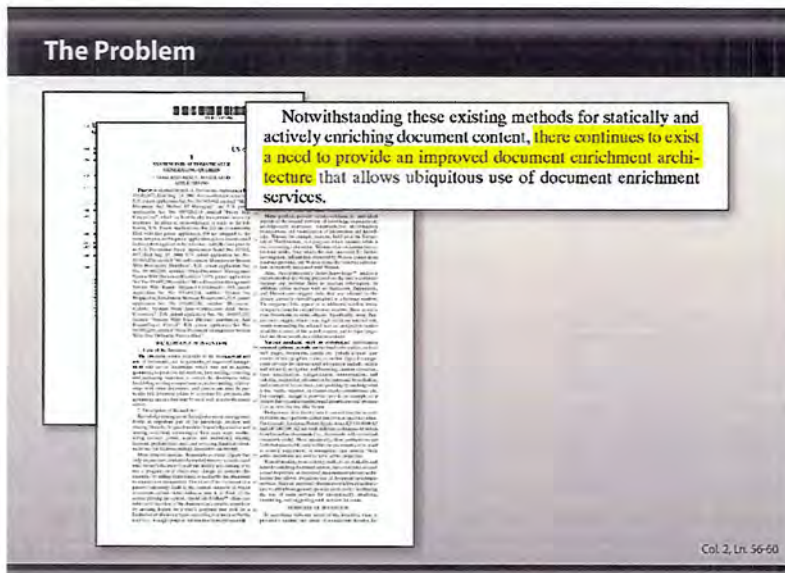
There is one patent at issue in this case: U.S. Patent Number 6,778,979.

The patent claims a method for automatically generating a query from selected document content.

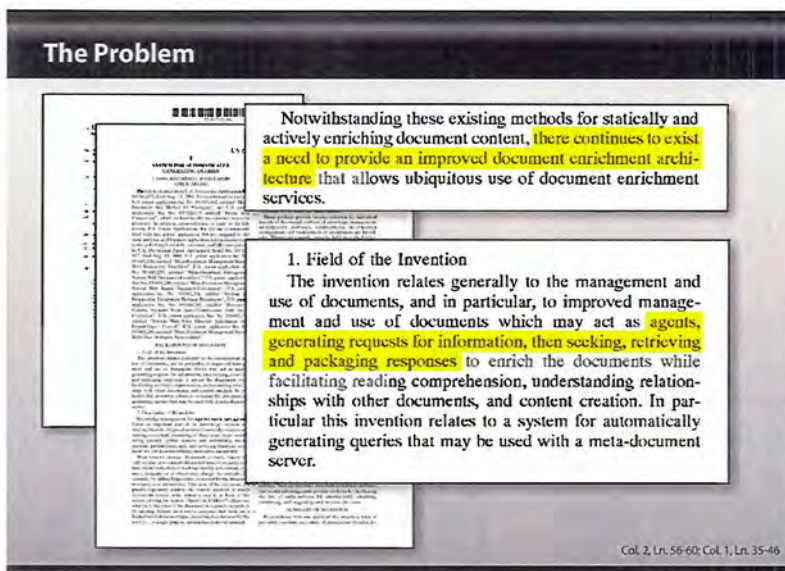


The '979 patent, issued in August 2004, relates to a system for dynamically obtaining additional information related to documents. According to the patent, most contemporary document management systems treated documents as “static objects” and “passive repositories” of information.

The Problem



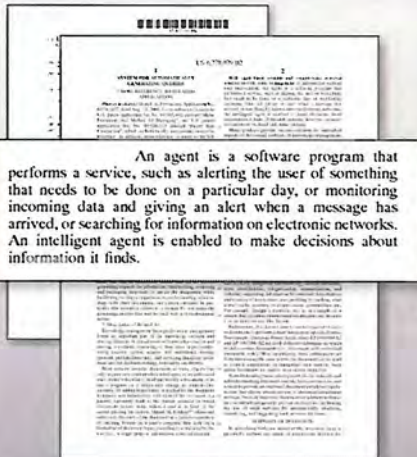
According to the patent, there existed a need for an “improved document enrichment architecture” that allowed for documents to be enriched with additional content without user intervention.



This functionality is added through the use of software agents which generate requests for additional information based on the content of the document.

Existing Solutions

Existing Solutions



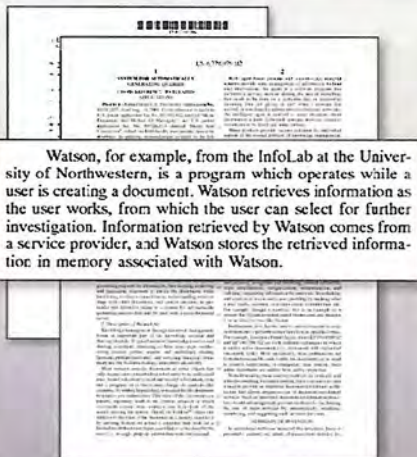
An agent is a software program that performs a service, such as alerting the user of something that needs to be done on a particular day, or monitoring incoming data and giving an alert when a message has arrived, or searching for information on electronic networks. An intelligent agent is enabled to make decisions about information it finds.

“An agent is a software program that performs a service...”

Col. 2, Ln. 3-9

Agent-based systems are software programs that perform certain tasks based on instructions or information, acting as the user’s “agents.” As the specification acknowledges, prior art agent-based systems were capable of providing document enrichment services.

Existing Solutions



Watson, for example, from the InfoLab at the University of Northwestern, is a program which operates while a user is creating a document. Watson retrieves information as the user works, from which the user can select for further investigation. Information retrieved by Watson comes from a service provider, and Watson stores the retrieved information in memory associated with Watson.

“An agent is a software program that performs a service...”

- Watson (Northwestern University InfoLab) monitors the user’s activity as he creates a document and retrieves relevant information from online service providers based on the user’s choices

Col. 2, Ln. 14-20

For example, an agent could be programmed to search the Internet for information related to words or phrases that a user is typing into a word processor, and present that information to the user. The Watson system from Northwestern’s InfoLab,

Existing Solutions



Also, Autonomy.com’s ActiveKnowledge™ analyzes documents that are being prepared on the user’s computer desktop and provides links to relevant information.

“An agent is a software program that performs a service...”

- Watson (Northwestern University InfoLab) monitors the user’s activity as he creates a document and retrieves relevant information from online service providers based on the user’s choices
- Autonomy.com’s ActiveKnowledge system analyzed documents created by the user to provide links to relevant information on the Web

Col. 2, Ln. 21-23

and the ActiveKnowledge system from Autonomy.com are two such solutions discussed in the patent.

Existing Solutions

Existing Solutions

in addition, online services such as Alexa.com, Zapper.com, and Flyswat.com suggest links that are relevant to the content currently viewed highlighted in a browser window. The suggested links appear in an additional window inside or separate from the current browser window. These services treat documents as static objects. Specifically, using Zapper.com's engine, when a user right clicks on selected text, words surrounding the selected text are analyzed to understand the context of the search request, and to reject pages that use those words in a different context.

- **Auto-link services: Zapper.com**
Provided suggested links to related documents based on the user's browsing activity

Col. 2, Ln. 23-33

The patent also lists other existing solutions to the asserted problem of document enrichment and management. One approach employed by products such as Alexa, Zapper, and Flyswat is to scan the content of the document, locate key words or phrases, and provide additional information to the user on those phrases if requested.

Existing Solutions

in addition, online services such as Zapper.com suggest links that are relevant to the content currently viewed highlighted in a browser window. The suggested links appear in an additional window inside or separate from the current browser window. These services treat documents as static objects. Specifically, using Zapper.com's engine, when a user right clicks on selected text, words surrounding the selected text are analyzed to understand the context of the search request, and to reject pages that use those words in a different context.

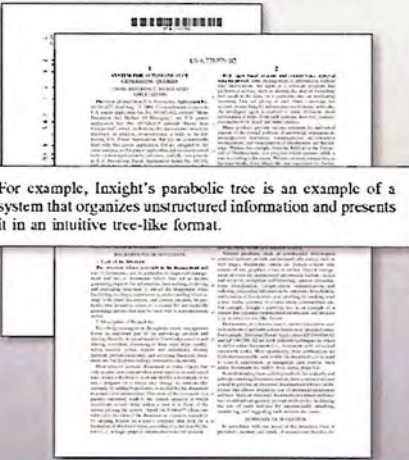
- **Auto-link services: Zapper.com**
Provided suggested links to related documents based on the user's browsing activity

Col. 2, Ln. 23-33

For example, Zapper might highlight text phrases like "Vikings," "Michael Jordan," or "Las Vegas" for the user. Should the user right click on the highlighted text, Zapper submits a query to a search engine for that text, then uses the words surrounding the text to filter the results.

Existing Solutions

Existing Solutions



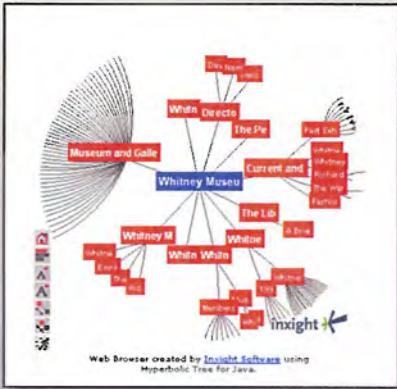
For example, Inxight's parabolic tree is an example of a system that organizes unstructured information and presents it in an intuitive tree-like format.

- **Auto-link services: Zapper.com**
Provided suggested links to related documents based on the user's browsing activity
- **Categorization services: Inxight**
Organized large amounts of content into a tree of nodes connected by links, allowing for easy navigation between related documents

Col. 2, Ln. 44-46

Other solutions, such as Inxight's Parabolic Tree, automatically categorize the links present on a document and present them to the user in a structured hierarchy of nodes or categories.

Existing Solutions



Web browser created by Inxight Software using Hyperbolic Tree for Java.

- **Auto-link services: Zapper.com**
Provided suggested links to related documents based on the user's browsing activity
- **Categorization services: Inxight**
Organized large amounts of content into a tree of nodes connected by links, allowing for easy navigation between related documents

Existing Solutions

Existing Solutions

Furthermore, it is known how to embed executable code in documents to perform certain functions at specified times. For example, European Patent Applications EP 0986010 A2 and EP 1087306 A2 set forth different techniques in which to define active documents (i.e., documents with embedded executable code). More specifically, these publications set forth that executable code within the document can be used to control, supplement, or manipulate their content. Such active documents are said to have active properties.

- **Auto-link services: Zapper.com**
Provided suggested links to related documents based on the user's browsing activity
- **Categorization services: Inxight**
Organized large amounts of content into a tree of nodes connected by links, allowing for easy navigation between related documents
- **Active documents: EP 0986010 A2/EP 1087306 A2**
Disclosed techniques for embedding executable code within a document to control, supplement, or manipulate the content of the document

Col. 2, Ln. 47-55

Active documents were also known in the art.

Existing Solutions

(57) A document management system is provided which organizes, stores and retrieves documents in accordance with document properties. A property attachment mechanism allows a user to define and attach static properties and/or active properties to a document. The active properties include executable code which control the behavior of the document contents. Upon transferring a document to another user, system, or environment, the document management system combines the document content and properties as a self-contained document which can interpret and manipulate its own contents. In this manner, a receiving user does not require additional applications in order to manipulate the document contents into a usable format. The self-contained document interprets and manipulates itself using its active properties to provide a useful document to the receiving user.

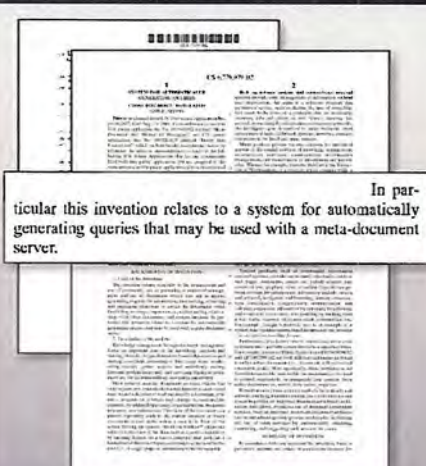
- **Auto-link services: Zapper.com**
Provided suggested links to related documents based on the user's browsing activity
- **Categorization services: Inxight**
Organized large amounts of content into a tree of nodes connected by links, allowing for easy navigation between related documents
- **Active documents: EP 0986010 A2/EP 1087306 A2**
Disclosed techniques for embedding executable code within a document to control, supplement, or manipulate the content of the document

EO 0986010

For example, the '010 and '306 European patent applications disclosed techniques for embedding executable software code within documents to manipulate the content of the documents based on a user's preferences.

The Claimed Invention

The Claimed Invention



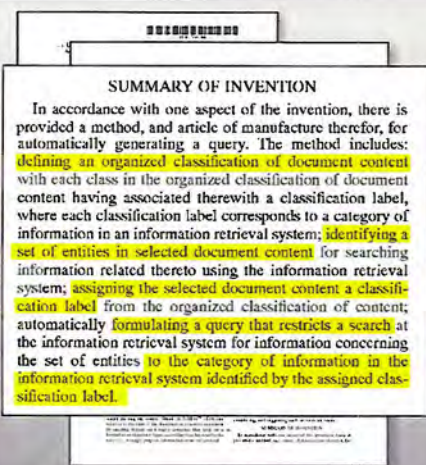
In particular this invention relates to a system for automatically generating queries that may be used with a meta-document server.

The claims address a specific facet of document enrichment: Constructing a search query based on document content

Col. 1, Ln. 43-46

While the specification of the '979 patent describes a number of aspects to enriching document content, the claims all relate to a "system for automatically generating queries that may be used with a meta-document server" such as a search engine.

The Claimed Invention



SUMMARY OF INVENTION

In accordance with one aspect of the invention, there is provided a method, and article of manufacture therefor, for automatically generating a query. The method includes: **defining an organized classification of document content** with each class in the organized classification of document content having associated therewith a classification label, where each classification label corresponds to a category of information in an information retrieval system; **identifying a set of entities in selected document content** for searching information related thereto using the information retrieval system; **assigning the selected document content a classification label** from the organized classification of content; **automatically formulating a query that restricts 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.

The claims address a specific facet of document enrichment: Constructing a search query based on document content

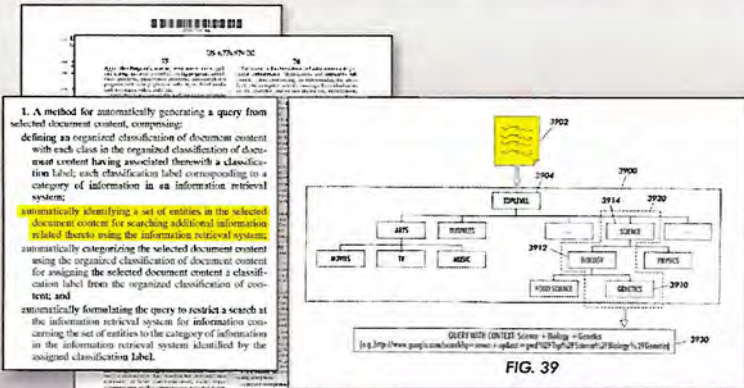
- Defining an organized classification of document content
- Identifying a set of entities in selected document content
- Assigning the selected document content a classification label
- Restricting searches for the entities to the category that corresponds to the classification label

Col. 2, Ln. 64 - Col. 3, Ln. 15

Specifically, the claims disclose a method of defining an organized classification of document content, identifying a set of entities or words within the document for performing additional searches, assigning the selected document content a classification label from the organized classification of document content, and automatically formulating a query to restrict a search at the information retrieval system concerning the identified words or entities to the corresponding category of information.

The Claimed Invention

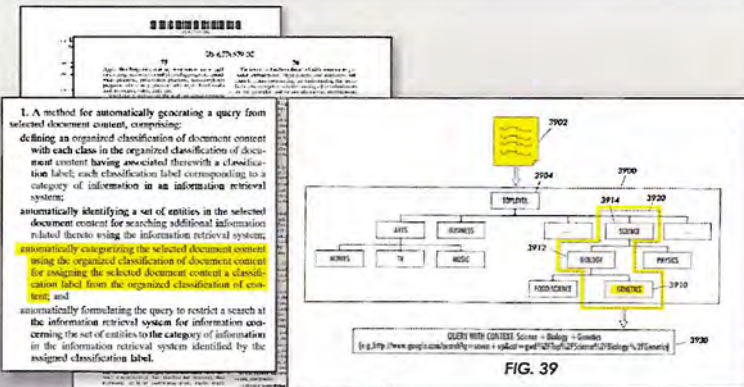
The Claimed Invention



Col. 76, Ln. 10-31

In the second step, a set of entities – words, phrases, or other portions of the document – is identified in the content of a document. Each word or phrase represents a concept about which a user may desire additional information.

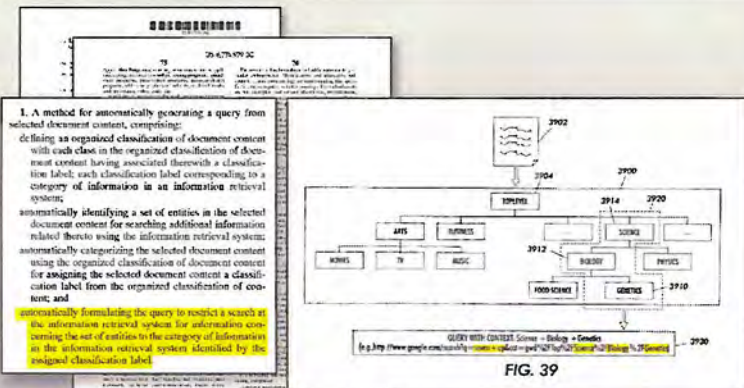
The Claimed Invention



Col. 76, Ln. 10-31

In the third step, selected document content is analyzed using the classification scheme and is assigned one classification label from the organized classification defined in the first step.

The Claimed Invention



Col. 76, Ln. 10-31

In the last step, a query is formulated to restrict a search to the category in the information retrieval system that corresponds to the selected document content's classification label.

Step 1: Defining an Organized Classification

Step 1: Defining an Organized Classification

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.

Col. 49, Ln. 18-30

Before the system can classify document content and use the classification to restrict a search, it must first define the classification scheme to be used.

Step 1: Defining an Organized Classification

Example in the patent: A hierarchical ontology such as the Open Directory Project (DMOZ)

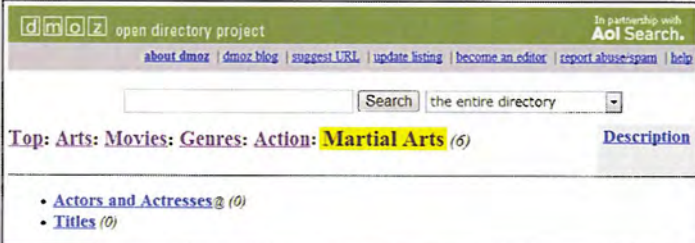
d m o z open directory project

An example of such a scheme given in the patent is the Open Directory Project, a publicly available content directory of links to World Wide Web sites also known as DMOZ. DMOZ uses a hierarchical ontology for organizing site listings by subject matter.

Step 1: Defining an Organized Classification

Step 1: Defining an Organized Classification

DMOZ hierarchical ontology




The screenshot shows the DMOZ directory interface. At the top, it says "DMOZ open directory project" and "In partnership with AOL Search." Below that are navigation links: "about dmoz", "dmoz blog", "suggest URL", "update listing", "become an editor", "report abuse/spam", and "help". There is a search bar with a "Search" button and a dropdown menu set to "the entire directory". The breadcrumb trail is "Top: Arts: Movies: Genres: Action: **Martial Arts** (6)", with "Martial Arts" highlighted in yellow. A "Description" link is to the right. Below the breadcrumb trail, there are two links: "Actors and Actresses @ (0)" and "Titles (0)".

For example, a document could be classified in the category "Martial Arts," which is itself a subcategory of the category

Step 1: Defining an Organized Classification

DMOZ hierarchical ontology




The screenshot shows the DMOZ directory interface. The breadcrumb trail is "Top: Arts: Movies: Genres: **Action: Martial Arts** (6)", with "Action: Martial Arts" highlighted in yellow. The "Description" link is to the right. Below the breadcrumb trail, there are two links: "Actors and Actresses @ (0)" and "Titles (0)".

"Action," which is a subcategory of

Step 1: Defining an Organized Classification

DMOZ hierarchical ontology



The screenshot shows the DMOZ directory interface. The breadcrumb trail is "Top: Arts: Movies: **Genres: Action: Martial Arts** (6)", with "Genres: Action: Martial Arts" highlighted in yellow. The "Description" link is to the right. Below the breadcrumb trail, there are two links: "Actors and Actresses @ (0)" and "Titles (0)".

"Genres," which is itself a subcategory of

Step 1: Defining an Organized Classification

Step 1: Defining an Organized Classification

DMOZ hierarchical ontology



The screenshot shows the DMOZ website interface. At the top, it says "dmoz open directory project" and "In partnership with AOL Search." Below that are navigation links: "about dmoz", "dmoz blog", "suggest URL", "update listing", "become an editor", "report abuse/spam", and "help". There is a search bar with a "Search" button and a dropdown menu set to "the entire directory". The main navigation bar shows "Top: Arts: **Movies**: Genres: Action: Martial Arts (6) Description". Below this, there are two sub-categories: "Actors and Actresses@ (0)" and "Titles (0)".

"Movies," which, finally, is a subcategory of the category

Step 1: Defining an Organized Classification

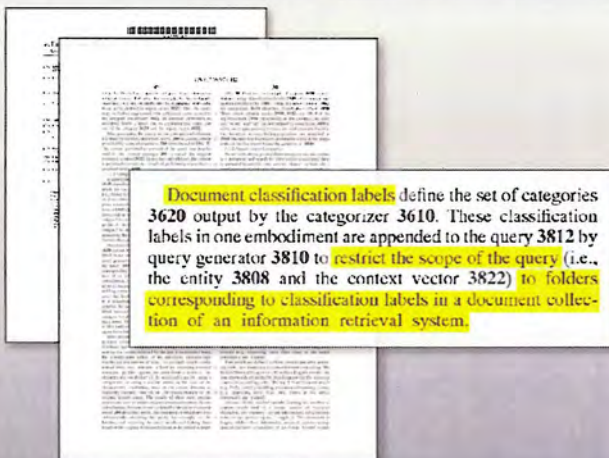
DMOZ hierarchical ontology



The screenshot shows the DMOZ website interface, similar to the previous one. The main navigation bar now shows "Top: **Arts**: Movies: Genres: Action: Martial Arts (6) Description". The sub-categories "Actors and Actresses@ (0)" and "Titles (0)" remain the same.

"Arts."

Step 1: Defining an Organized Classification

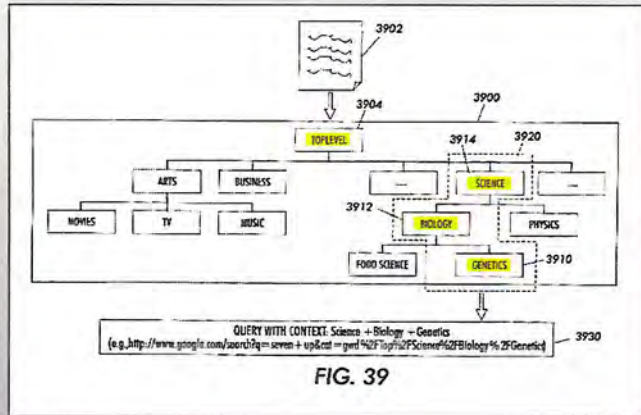


The screenshot shows a document with a highlighted text box. The text in the box reads: "Document classification labels define the set of categories 3620 output by the categorizer 3610. These classification labels in one embodiment are appended to the query 3812 by query generator 3810 to restrict the 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."

Importantly, the categories that are defined in this step for the purpose of document classification need to correspond to categories in the information retrieval system that will ultimately be searched. As we will see later on, this allows the claimed process to restrict a search at the information retrieval system to the assigned category.

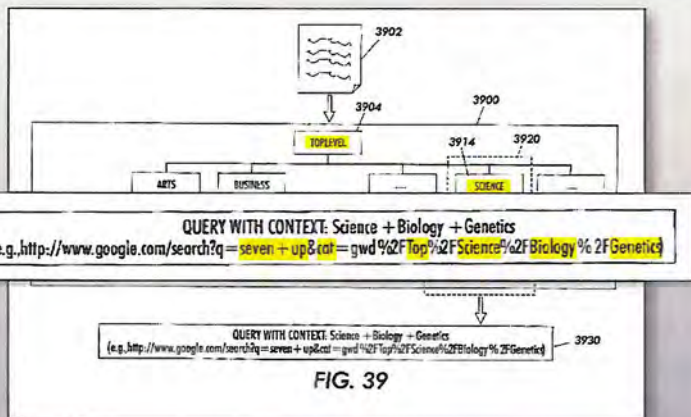
Step 1: Defining an Organized Classification

Step 1: Defining an Organized Classification



The example illustrated in figure 39 of the patent is similar to DMOZ – a nested hierarchy of categories (or “nodes”) that becomes increasingly specific lower in the tree structure: Genetics is a subcategory of biology, which is itself a subcategory of science, which itself is a subcategory of the “toplevel” or “root” node.

Step 1: Defining an Organized Classification



Note how the category is used to formulate a query to restrict a search. Here, the search for “seven up” has been restricted to the category “Top/Science/Biology/Genetics.” This ensures that a search concerning “seven up” is done for documents in the category “genetics” rather than in a category for soft drinks.

Step 2: Identifying Entities

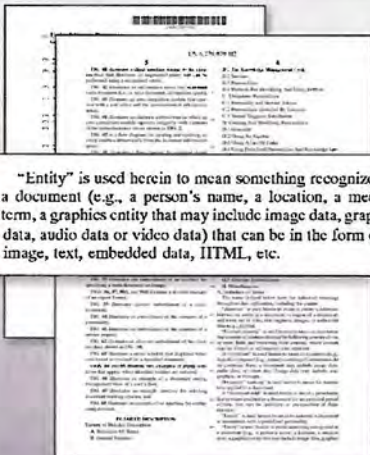
Step 2: Identifying Entities



Entity: "something recognized in a document (e.g., a person's name, a location, a medical term, a graphics entity that may include image data, graphics data, audio data or video data) that can be in the form of an image, text, embedded data, HTML, etc."

After the claimed system has its defined organized classification, it can formulate queries concerning the subject matter of document content to restrict a search. The system scans document content for "entities," which are meant to represent concepts that might interest the user.

Step 2: Identifying Entities



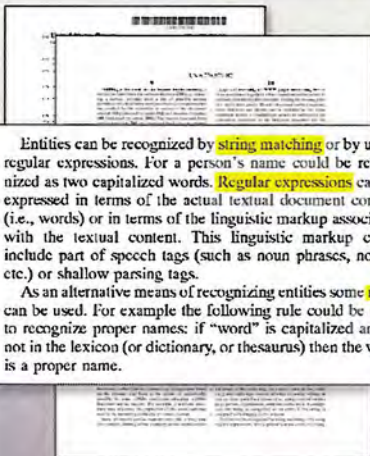
"Entity" is used herein to mean something recognized in a document (e.g., a person's name, a location, a medical term, a graphics entity that may include image data, graphics data, audio data or video data) that can be in the form of an image, text, embedded data, HTML, etc.

Entity: "something recognized in a document (e.g., a person's name, a location, a medical term, a graphics entity that may include image data, graphics data, audio data or video data) that can be in the form of an image, text, embedded data, HTML, etc."

"Entity" is defined in the specification as "something recognized in a document" and may include names, places, or other proper nouns. For example, an entity may be the name of a person like "Michael Jordon" or "Ben Franklin," the name of a city like "Paris" or "Philadelphia," or the name of a product like "Toyota Prius" or "Furby."

Col. 6, Ln. 65 - Col. 7, Ln. 2

Step 2: Identifying Entities



Entities can be recognized by **string matching** or by using regular expressions. For a person's name could be recognized as two capitalized words. **Regular expressions** can be expressed in terms of the actual textual document content (i.e., words) or in terms of the linguistic markup associated with the textual content. This linguistic markup could include part of speech tags (such as noun phrases, nouns, etc.) or shallow parsing tags.

As an alternative means of recognizing entities some **rules** can be used. For example the following rule could be used to recognize proper names: if "word" is capitalized and is not in the lexicon (or dictionary, or thesaurus) then the word is a proper name.

Entity: "something recognized in a document (e.g., a person's name, a location, a medical term, a graphics entity that may include image data, graphics data, audio data or video data) that can be in the form of an image, text, embedded data, HTML, etc."

Entities are identified in document content and used to search for related information

The patent discusses various ways of identifying entities in a document, including string matching, regular expressions, and rules, all of which were known in the art.

Col. 10, Ln. 66 - Col. 11, Ln. 11

Step 2: Example

Step 2: Example

Eagles finalize coaching staff

AP Associated Press

Share

0

Email

Print

- Tue Feb 8, 3:02 pm ET

PHILADELPHIA – Andy Reid and the Philadelphia Eagles completed their staff on Tuesday by adding two coaches and promoting four others.

Johnnie Lynn was hired to be the secondary/cornerbacks coach, and Bobby April, Jr. was named defensive quality control coach.

David Culley was promoted to senior offensive assistant/wide receivers. James Urban becomes assistant offensive coordinator. Doug Pederson is the new quarterbacks coach. And Duce Staley will serve as special teams quality control coach.

Lynn joins the Eagles after a five-year stint in San Francisco, working as the 49ers secondary coach. April, the son of Philadelphia's special teams coordinator Bobby April, spent 2010 as special teams coordinator and safeties coach at Nicholls State.

Culley has been the wide receivers coach since 1999. Urban spent the past two seasons as quarterbacks coach. Pederson worked as the team's offensive quality control coach from 2009-10. And Staley was a coaching intern last season.

For example, suppose a user were viewing a Web page containing a news article about the Philadelphia Eagles. The system might scan the article looking for entities, that is, words or phrases, that represent topics toward which it can direct searches.

Step 2: Example

Eagles finalize coaching staff

AP Associated Press

Share

0

Email

Print

- Tue Feb 8, 3:02 pm ET

PHILADELPHIA – Andy Reid and the Philadelphia Eagles completed their staff on Tuesday by adding two coaches and promoting four others.

Johnnie Lynn was hired to be the secondary/cornerbacks coach, and Bobby April, Jr. was named defensive quality control coach.

David Culley was promoted to senior offensive assistant/wide receivers. James Urban becomes assistant offensive coordinator. Doug Pederson is the new quarterbacks coach. And Duce Staley will serve as special teams quality control coach.

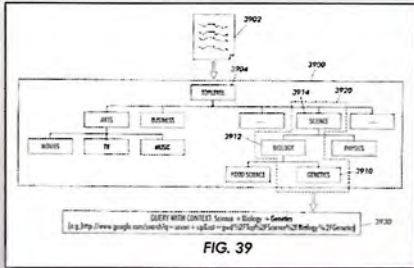
Lynn joins the Eagles after a five-year stint in San Francisco, working as the 49ers secondary coach. April, the son of Philadelphia's special teams coordinator Bobby April, spent 2010 as special teams coordinator and safeties coach at Nicholls State.

Culley has been the wide receivers coach since 1999. Urban spent the past two seasons as quarterbacks coach. Pederson worked as the team's offensive quality control coach from 2009-10. And Staley was a coaching intern last season.

In this case, the system might identify "Andy Reid," "Philadelphia Eagles," "Johnnie Lynn," and "Bobby April, Jr." as entities, among others.

Step 3: Categorizing to Assign a Classification Label to Document Content

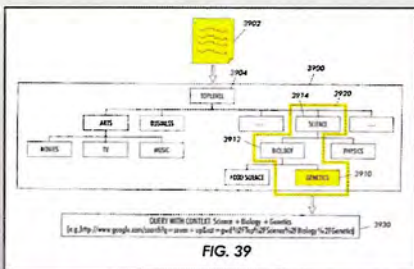
Step 3: Categorizing to Assign a Classification Label to Document Content



The system also assigns document content a classification label corresponding to a category from the organized classification

After identifying entities within the document content, the system assigns document content a classification label that corresponds to a category from the organized classification discussed earlier.

Step 3: Categorizing to Assign a Classification Label to Document Content



The system also assigns document content a classification label corresponding to a category from the organized classification

In the example given in figure 39, the document is given the classification label belonging to the subcategory "genetics."

Step 3: Categorizing to Assign a Classification Label to Document Content

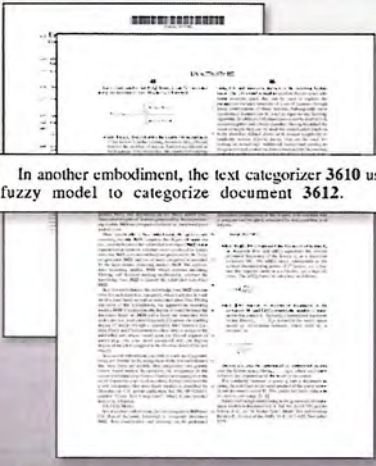


- The system also assigns document content a classification label corresponding to a category from the organized classification
- The classification can be accomplished in a number of different ways, all known in the art

The classification can be accomplished by various means known in the art at the time of the patent.

Step 3: Categorizing to Assign a Classification Label to Document Content

Step 3: Categorizing to Assign a Classification Label to Document Content



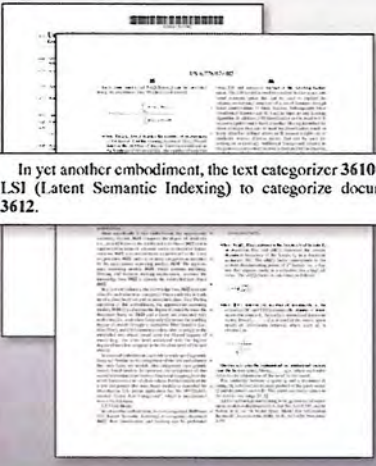
In another embodiment, the text categorizer 3610 uses a fuzzy model to categorize document 3612.

- The system also assigns document content a classification label corresponding to a category from the organized classification
- The classification can be accomplished in a number of different ways, all known in the art

Col. 45, Ln. 18-19

The patent discusses various such methods, for example, fuzzy models,

Step 3: Categorizing to Assign a Classification Label to Document Content




In yet another embodiment, the text categorizer 3610 uses LSI (Latent Semantic Indexing) to categorize document 3612.

- The system also assigns document content a classification label corresponding to a category from the organized classification
- The classification can be accomplished in a number of different ways, all known in the art

Col. 45, Ln. 65-67

latent semantic indexing,

Step 3: Categorizing to Assign a Classification Label to Document Content



In yet a further embodiment, the text categorizer 3610 uses a vector space model to categorize document 3612.

- The system also assigns document content a classification label corresponding to a category from the organized classification
- The classification can be accomplished in a number of different ways, all known in the art

Col. 46, Ln. 18-19

and a vector space model.

Step 3: Example

Step 3: Example

Eagles finalize coaching staff

Associated Press

PHILADELPHIA – Andy Reid and the Philadelphia Eagles completed their staff on Tuesday by adding two coaches and promoting four others.

Johnnie Lynn was hired to be the secondary/cornerbacks coach, and Bobby April, Jr., was named defensive quality control coach.

David Culley was promoted to senior offensive assistant/wide receivers. James Urban becomes assistant offensive coordinator. Doug Pederson is the new quarterbacks coach. And Duce Staley will serve as special teams quality control coach.

Lynn joins the Eagles after a five-year stint in San Francisco, working as the 49ers secondary coach. April, the son of Philadelphia's special teams coordinator Bobby April, spent 2010 as special teams coordinator and safeties coach at Nicholls State.

Culley has been the wide receivers coach since 1999. Urban spent the past two seasons as quarterbacks coach. Pederson worked as the team's offensive quality control coach from 2009-10. And Staley was a coaching intern last season.

Going back to the example, the system would analyze content, then try to match the Eagles news article with one of the categories in the DMOZ hierarchy.

Step 3: Example

Eagles finalize coaching staff

Associated Press


PHILADELPHIA – Andy Reid and the Philadelphia Eagles completed their staff on Tuesday by adding two coaches and promoting four others.

Johnnie Lynn was hired to be the secondary/cornerbacks coach, and Bobby April, Jr., was named defensive quality control coach.

David Culley was promoted to senior offensive assistant/wide receivers. James Urban becomes assistant offensive coordinator. Doug Pederson is the new quarterbacks coach. And Duce Staley will serve as special teams quality control coach.



Lynn joins the Eagles after a five-year stint in San Francisco, working as the 49ers secondary coach. April, the son of Philadelphia's special teams coordinator Bobby April, spent 2010 as special teams coordinator and safeties coach at Nicholls State.

Culley has been the wide receivers coach since 1999. Urban spent the past two seasons as quarterbacks coach. Pederson worked as the team's offensive quality control coach from 2009-10. And Staley was a coaching intern last season.



Categorizer

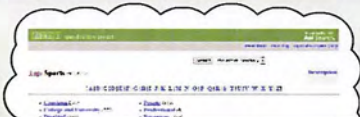
Step 3: Example




Categorizer

Step 3: Example

Step 3: Example



Top: Sports (92,472)



Categorizer

In this case, the categorizer might determine that the news article falls into the “Top,” “Sports,”

Step 3: Example



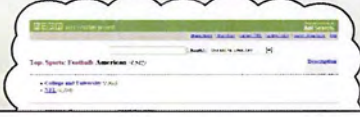
Top: Sports: Football (6,853)



Categorizer

“Football,”

Step 3: Example



Top: Sports: Football: American (4,541)



Categorizer

“American,”