(54) Title: OPTIMIZATION OF FACT EXTRACTION USING A MULTI-STAGE APPROACH

(57) Abstract: Facts are extracted from electronic documents by recognizing factual descriptions using a fact-word table to match to words of the electronic documents. The words of those factual descriptions may be tagged with the appropriate part of speech. More detailed analysis is then performed on those factual descriptions, rather than on the entire electronic document, and particularly to the text in the neighborhood of the fact-word matches. The analysis may involve identifying the linguistic constituents of each phrase and determining the role as either subject or object. Exclusion rules may be applied to eliminate those phrases unlikely to be part of facts, the exclusion rules being based in part on the linguistic constituents. Scoring rules may be applied to remaining phrases, and for those phrases having a score in excess of a threshold, the corresponding sentence part, whole sentence, paragraph, or other document portion may be presented as representing one or more facts.
OPTIMIZATION OF FACT EXTRACTION USING A MULTI-STAGE APPROACH

BACKGROUND

Electronic documents may contain a mixture of facts and opinions. At times, a reader may only be interested in facts, or may wish to have the facts be identified. For example, a user performing an on-line search for information may wish to obtain facts about a particular subject as quickly and efficiently as possible. However, presenting a list of web pages or other electronic documents that are related to the search terms used require the user to individually examine each web page or other electronic document and distinguish the facts from the opinions or subjective information.

Attempts have been made to perform fact extraction. However, accurate fact extraction can be a slow and inefficient process even for high-speed server computers. Such fact extraction attempts generally apply a linguistic analysis to the entire contents of the electronic document to extract those facts that it may contain. When applying fact extraction to hundreds or thousands of electronic documents, the amount of time needed to achieve a result may be unacceptable.

SUMMARY

Embodiments provide optimization of fact extraction by using a multi-stage approach. The electronic documents are scanned to find factual descriptions that are likely to contain facts by using a fact-word table to match terms within sentences of the electronic documents to obtain a set of factual descriptions. Further analysis may then be performed, including determining linguistic constituents, e.g., syntactic constituents and/or semantics, in the neighborhood of that set of factual descriptions rather than on the entire document. Accordingly, time is saved by avoiding a complex lexical and syntactic analysis of the entire document for every electronic document of interest.

This Summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This Summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used as an aid in determining the scope of the claimed subject matter.
BRIEF DESCRIPTION OF DRAWINGS

FIG. 1 shows an example of a computer system for implementing embodiments.

FIG. 2 shows an example of an operational flow of a search involving the presentation of facts that have been extracted prior to the search.

FIG. 3 shows an example of an operational flow of a search involving the presentation of facts that have been extracted during the search.

FIG. 4 shows an example of an operational flow of the multiple steps of fact extraction.

FIG. 5 shows an example of a more detailed operational flow of the multiple steps of fact extraction.

FIG. 6 shows an example of a screen display providing search results that include the presentation of facts obtained from electronic documents discovered by the search.

DETAILED DESCRIPTION

Embodiments provide for fact extraction using multiple stages to avoid performing complex analyses of the entire documents of interest. Factual descriptions of the documents are recognized in relation to a fact-word table in an initial stage. These factual descriptions may be tagged with their parts of speech, either noun or verb. Then more detailed analyses may be done in a subsequent stage over those factual descriptions to thereby avoid such detailed analyses over the entire documents of interest. The linguistic constituents for each factual description may be determined and then exclusions and scores may be used to eliminate factual descriptions that are less likely to be facts. The factual descriptions remaining after the exclusions and scoring may then be presented as fact.

FIG. 1 shows an example of a computer system 100 that provides an operating environment for the embodiments. The computer system 100 as shown may be a standard, general-purpose programmable computer system 100 including a processor 102 as well as various components including mass storage 112, memory 104, a display adapter 10S, and one or more input devices 110 such as a keyboard, keypad, mouse, and the like. The processor 102 communicates with each of the components through a data signaling bus 106. The computer system 100 may also include a network interface
124, such as a wired or wireless connection, that allows the computer system 100 to communicate with other computer systems via data networks. The computer system 100 may alternatively be a hard-wired, application specific device that implements one or more of the embodiments.

In the example, of FIG. 1, the processor 102 implements instructions stored in the mass storage 112 in the form of an operating system 114. The operating system 114 of this example provides a foundation upon which various applications may be implemented to utilize the components of the computer system 100. The computer system 100 may implement a search engine 118 or similar application for finding electronic documents relevant to a particular situation. For example, the search engine 118 may receive search terms entered directly through input device 110 by a user of the computer system 100 or may receive search terms submitted by a user of a remote computer that are received via the network interface 122.

The search and/or fact extraction may occur in relation to one or more sets of electronic documents that contain textual information such as web pages, standard word processing documents, spreadsheets, and so forth. These electronic documents may be stored locally as electronic document set 116. These electronic documents may also be stored at a non-local location such as network-based storage 124 containing an electronic document set 126. Network-based storage 124 is representative of local network storage, on-line storage locations of the Internet, and so forth. The network-based storage 124 is accessible via the network interface 122.

Additionally, these embodiments provide logic for implementation by the processor 102 in order to extract the facts from the electronic documents 116, 126. A fact extraction tool 120 may be present on the local storage device 112, either as a component of the operating system 114, a component of the search engine 118 or other application, or as a stand-alone application capable of producing its own independent results. The logical operations performed by embodiments of the fact extraction tool 120 are discussed below in relation to PIGS. 2-5.

The computer system 100 of FIG. 1 may include a variety of computer readable media. Such computer readable media contains the instructions for operation of the computer system and for implementation of the embodiments discussed herein. Computer readable media can be any available media that can be accessed by computer 100 and includes both volatile and nonvolatile media, removable and non-removable
media. By way of example, and not limitation, computer readable media may comprise computer storage media and communication media.

Computer storage media includes volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by computer system 100.

Communication media typically embodies computer readable instructions, data structures, program modules or other data in a modulated data signal as a carrier wave or other transport mechanism and includes any information delivery media. The term "modulated data signal" means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media includes wired media such as a wired network or direct-wired connection, and wireless media such as acoustic, RF, infrared and other wireless media. Combinations of the any of the above should also be included within the scope of computer readable media.

FIG. 2 shows an example of logical operations performed by a search engine 118 in conjunction with the fact extraction tool 120. In this example, the fact extraction tool 120 is utilized prior to a search occurring in order to generate a library of facts present in the electronic documents to be searched. In this manner, there is no processing time required to extract the facts but instead those facts have already been extracted and are retrieved from a fact library on the basis of the search terms entered.

The logical operations begin at collection operation 202 where the collection of electronic documents is obtained or access is otherwise achieved. For example, the electronic documents to eventually be searched may be saved to local storage or may be acquired via on-line access. The fact extraction tool 120 then operates upon each one of those electronic documents to attempt to extract all of the facts that are present in the electronic documents. The fact extraction tool 120 may generate a library of facts that are stored in association with the corresponding electronic documents and are available
for access during future searches. For example, Table 1 shows such a library of associations.

<table>
<thead>
<tr>
<th>Electronic Document</th>
<th>Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.sample1.com">www.sample1.com</a></td>
<td>Fact A</td>
</tr>
<tr>
<td></td>
<td>Fact B</td>
</tr>
<tr>
<td></td>
<td>Fact C</td>
</tr>
<tr>
<td><a href="http://www.sample2.com">www.sample2.com</a></td>
<td>Fact AA</td>
</tr>
<tr>
<td></td>
<td>Fact BB</td>
</tr>
<tr>
<td></td>
<td>Fact CC</td>
</tr>
<tr>
<td><a href="http://www.sample3.com">www.sample3.com</a></td>
<td>Fact AAA</td>
</tr>
</tbody>
</table>

Table 1

Continuing with the operational flow of FIG. 2, a user wishing to do a search to find relevant electronic documents, and particularly to find relevant facts from those electronic documents, enters a search term into the search engine 118 at term operation 206. In this example, the search engine 118 then searches through the electronic documents for the search terms and finds matching documents at document operation 208. The search engine also finds the previously extracted facts which match the search terms from those matching electronic documents and then displays the relevant documents or a link thereto along with the relevant facts at display operation 210. For example, a search term may be found in www.sample1.com and the search term may also be found to match Fact A and Fact B such that a link to www.sample1.com is displayed along with Fact A and Fact B. Thus, the user is quickly provided with facts related to the search terms that were entered. An example of such a screen display is discussed below in relation to FIG. 6.

Of course, as an alternative the search may be for previously extracted facts only, rather than for the electronic documents themselves. Furthermore, in certain circumstances the previously extracted facts may match the search terms regardless of whether the electronic documents containing the facts match the search terms.

FIG. 3 shows another example of logical operations performed by a search engine 118 in conjunction with the fact extraction tool 120. In this example, the fact extraction tool 120 is utilized during a search in order to discover facts present in the
electronic documents as they are being found by the search. In this manner, there is no need for pre-search fact extraction and no need for storage of a library of facts. In such a scenario, the fact extraction tool may only scan snippets or summaries of the document to provide very fast results, or the entire document may also be scanned to extract all potential facts.

The logical operations begin at term operation 302 where a user enters a search term into the search engine 118. In this example, the search engine 118 then searches through the electronic documents for the search terms and finds matching documents at document operation 304. The extraction tool 120 is then employed at extraction operation 306 in order to analyze the electronic documents that have been found by the search in order to extract facts from those documents that are relevant to the search terms. The result of extraction operation 306 may produce a temporary set of associations between electronic documents and facts as shown in Table 1, which may then be placed in longer term storage in anticipation searches for those search terms occurring in the future. The search engine then displays the relevant documents or a link thereto along with the relevant facts returned by the fact extraction tool 120 at extraction operation 306 at display operation 308.

FIG. 4 shows the multi-stage approach utilized by embodiments of the fact extraction tool 120. Initially, the fact extraction tool 120 attempts to recognize a set of factual descriptions from the electronic documents of interest at recognition operation 402. Here, the goal is to find those descriptions in the text that are likely to be facts based on finding matches to a fact-word table discussed in more detail below with reference to FIG. 5. By performing a quick matching process, much of the electronic document that should be ignored when finding facts can be eliminated from further fact extraction processing thereby increasing the efficiency of the subsequent stage(s) that are employed to increase accuracy.

After having identified a set of factual descriptions for a document being analyzed, fact extraction is then performed on that set of factual descriptions at extraction operation 404. Here, more detailed analyses are performed only on the set of factual descriptions, as opposed to the whole document, so that satisfactory efficiency is maintained while adequate accuracy is achieved. The analyses of extraction operation involve decision making based on a determination of linguistic constituents of the
factual descriptions. Such linguistic constituents may include the syntactic constituents, the semantics, and so forth.

FIG. 5 shows an example of details of the recognition and extraction operations of FIG. 4. The logical operations begin at scanning operation 502 where the fact extraction tool 120 scans the electronic document to find words or phrases matching those of a fact-word table. A fact-word table is a list of words or phrases that are known to likely be used when expressing a fact as opposed to an opinion for example. Table 2 shows a brief example. Note that to provide optimal processing performance, the words of the table may be associated with the most appropriate part of speech (POS) tag which is discussed below in relation to tag operation 504.

<table>
<thead>
<tr>
<th>Fact-Word List</th>
<th>POS Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word / Phrase 1</td>
<td>POS Tag</td>
</tr>
<tr>
<td>Word / Phrase 2</td>
<td>POS Tag</td>
</tr>
<tr>
<td>Word / Phrase N</td>
<td>POS Tag</td>
</tr>
</tbody>
</table>

Table 2

Research has been done to determine words that are suggestive of facts rather than opinions. For example, the class of words that introduce facts can be derived using research and work on the classification of verbs and their lexical functions. Two relevant papers that may be used as a material to do so include:


Thus, on the basis of such research, the fact-word list as shown in Table 2 may be constructed to include those verbs or other words that are suggestive of a fact expression as opposed to a non-fact. For example, the terms "invented" or "hired" are suggestive of a fact expression whereas the terms "can be" or "complains" are not. A particular example of a fact-word list can be found in Appendix A located at the end of
this specification. This particular example is a non-exhaustive list of verbs that are fact-words that may be used to discover factual descriptions in electronic documents.

Either upon application of the fact-word table to an electronic document, or in parallel with the application of the fact-word table such as where the POS Tag is already associated with the words in the fact-word table, the parts of speech (POS) of each of the words of each factual description are tagged at tag operation 504. This tagging operation 504, which may occur in parallel with or subsequent to scan operation 502, may involve making disambiguating choices for words which have more than one POS tag, such as by favoring a noun tag over a verb tag since it is understood that syntactic phrases like noun phrases are known to be the entities involved in a factual event. Any unknown and non-pre-tagged words may default to nouns for this reason as well. As with nouns, adjectives may be favored over verbs (e.g., "planned" as an adjective over "planned" as a verb) as well such that words having both an adjective and verb tag will default to adjective because adjective are part of noun phrases which are known to be the entities involved in a factual event. When creating the associations of the POS Tags to the words of the fact-word table, such as when creating the table, these disambiguating choices may already be applied so that, for instance, "planned" is associated with an adjective POS Tag in the table and not a verb POS Tag.

Once the factual descriptions have been found and the words of the factual descriptions have been tagged with the POS, then the more complete analysis may be performed to improve the accuracy of the fact extraction without requiring that the entire document be subjected to this more complete processing. At identification operation 506, syntactic phrases like noun phrases and verb phrases are identified. The syntactic phrases are identified by utilizing conventional grammar rules and light linguistic analysis. Those syntactic phrases that are in the neighborhood, i.e., very local to the set of factual descriptions in a document are identified and if a factual description has no syntactic phrases associated to it, then the corresponding sentence may be eliminated from further consideration. Thus, by focusing on only those syntactic phrases that are in the neighborhood of the factual description, the process avoids looking at all the linguistic constituents of a whole sentence.

Furthermore, at identification operation 506, the linguistic constituents of the factual descriptions having the neighboring syntactic phrases are further determined by assessing the role a syntactic phrase plays within the corresponding sentence based on
the pattern identified in the factual description. Thus, it is determined from the word pattern of the factual description whether the syntactic phrase plays the role of subject or object within the sentence containing the current factual description being analyzed.

Once the linguistic constituents of the factual descriptions are determined, i.e., the syntactic phrases and their roles have been identified, exclusion rules may then be applied to those noun phrases of the factual descriptions to further eliminate those that are less likely to be an expression of fact at exclusion operation 508. The exclusion rules may be applicable on the basis of a syntactic phrase as an object, a syntactic phrase as a subject, or a syntactic phrase without regard to its role. Furthermore, in this particular embodiment, an exclusion rule being applied to individual words, to the syntactic phrases, or to the whole sentence lead to the same result, which is to exclude the whole sentence from being a factual description. An example of exclusion rules that may be applied is shown in Table 3.

<table>
<thead>
<tr>
<th>Exclusion Rules</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Object” has “opinion/biased” modifier</td>
<td>Rule out the sentence candidate</td>
</tr>
<tr>
<td>Sentence Filters:</td>
<td></td>
</tr>
<tr>
<td>- Initial word of sentence (e.g., pronouns)</td>
<td>Rule out the sentence candidate</td>
</tr>
<tr>
<td>- Punctuation: e.g. “?”</td>
<td></td>
</tr>
<tr>
<td>“Subject” is a definite - unless Proper name</td>
<td>Rule out the sentence candidate</td>
</tr>
<tr>
<td>Surrounding “Context” of the “Object”</td>
<td>Rule out the sentence candidate if the</td>
</tr>
<tr>
<td></td>
<td>surrounding context has a particular POS</td>
</tr>
<tr>
<td></td>
<td>that is not indicative of a fact (e.g., some</td>
</tr>
<tr>
<td></td>
<td>class of pronouns)</td>
</tr>
<tr>
<td>Stop words occur in the sentence</td>
<td>Rule out the sentence candidate</td>
</tr>
<tr>
<td>“Subject” of “Object” contain pronouns</td>
<td>Rule out Noun Phrase</td>
</tr>
</tbody>
</table>

Table 3
Either upon application of the exclusion rules, or in parallel with the application of the exclusion rules, scoring rules are applied at scoring operation 510. The scoring rules give a weight to both the subject and object noun phrases for each of various features, and a total score for the candidate factual description is the sum of the individual feature weights plus the certainty score of the matching fact-word. The individual feature weights may be positive, when indicative of a fact, and may be negative, when indicative of a non-fact. Examples of features and associated scoring rules are provided below in Table 4. The feature scores may be manually assigned using human judgment or may be automatically learned.

<table>
<thead>
<tr>
<th>Features</th>
<th>Scoring rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Certainty score of the matching pattern (fact-word, e.g., main verb)</td>
<td></td>
</tr>
<tr>
<td>Class of the Roles (i.e., subject or verb), e.g.: person, country, organization, etc.</td>
<td>Score per class</td>
</tr>
<tr>
<td>Main “subject” contains a Proper Name</td>
<td>Normal weight</td>
</tr>
<tr>
<td>“Object” length</td>
<td>Length score</td>
</tr>
<tr>
<td>“Subject” length</td>
<td>Length score</td>
</tr>
<tr>
<td>Sentence length</td>
<td>Length score</td>
</tr>
<tr>
<td>“Subject” appears at beginning of sentence — i.e., subject offset</td>
<td>Position score</td>
</tr>
<tr>
<td>“Object” has a modifier (adjective, adverbs)</td>
<td>Negative - Basic weight</td>
</tr>
<tr>
<td>“Object” is a definite (“the”)</td>
<td>Negative – Basic</td>
</tr>
<tr>
<td></td>
<td>Exclusive when ends copula sentence</td>
</tr>
</tbody>
</table>

Table 4

The total score for the factual description is then compared to a pre-defined threshold to determine whether the total score exceeds the threshold at query operation 512. If the threshold is not exceeded, then the corresponding factual description may be discarded. If the threshold is exceeded, then the factual description, the complete sentence, and/or the complete paragraph or other document portion may be presented as
a fact at presentation operation 514. This presentation may include displaying the fact, saving the fact to a library, and so forth.

In utilizing the scoring rules and threshold comparison, the weights assigned to the features and/or the threshold value may be manipulated without manipulating the whole approach to fact extraction. In this manner, the degree of accuracy of fact extraction and presentation can be controlled while the processing steps remain the same.

FIG. 6 shows an example screenshot 600 resulting from performing a search. Search terms have been entered in search field 602 to conduct the search. The search term has been matched to various web site links 604 available from the Internet. The user may visit the electronic documents in the normal fashion.

Additionally, facts 610, 612, and 614 about the search term are displayed in section 608. Accordingly, a user can quickly spot facts about the subject of the search without having to visit any of the electronic documents that have been found and without having to manually read and discern fact from opinion. In this particular example, the facts 610, 612, and 614 include hyperlinks that the user may select to give more information about the source of the fact and/or to show the context within which the facts were discovered (e.g., date of the fact associated, other facts, etc.).

It will be appreciated that screenshot 600 is merely one example of how the facts may be presented to the user. Rather than presenting them in a separate column as shown, they may be listed as sub-elements of the electronic document that they have been extracted from. Furthermore, as an alternative to or in addition to listing the facts on the search results page, the facts extracted from a particular electronic document may also be listed in a column or other location upon the user viewing the electronic document itself. Additionally, as an alternative to or in addition to separating the facts from the document for display, the facts may be highlighted within the electronic documents both in the list of documents 604 within the search results and within the complete electronic document when it is chosen for display. As yet another alternative, the facts may be displayed independently from search results, such as to display facts only with a selectable link to obtain the source documents, where only the extracted facts have been searched to thereby avoid the document search completely.

Additionally, it will be appreciated that the presentation of the extracted facts, such as that shown in screenshot 600, may be provided as a display to a local computer
implementing the search and fact extraction for a local user. Alternatively, the presentation of the extracted facts, such as that shown in screenshot 600, may be provided as a display to a remote computer that has requested that the local computer perform the search and fact extraction on its behalf, such as in the case of an Internet based search engine.

Accordingly, facts may be efficiently and accurately extracted from documents for presentation to users. Through the multi-stage approach, the efficiency is increased by avoiding detailed analysis of the entire documents as well as avoiding detailed analysis of the entire sentence where a factual description has been found. The accuracy is maintained by employing further analysis upon the factual descriptions that have been discovered in the document by the initial stage of processing.

While the invention has been particularly shown and described with reference to various embodiments thereof, it will be understood by those skilled in the art that various other changes in the form and details may be made therein without departing from the spirit and scope of the invention. For example, certain exclusion rules that are not specific to the linguistic constituents of a factual description, such as those based on punctuation of a sentence, may be applied when parsing for the factual description rather than later during the application of other exclusion rules.

Appendix A - Fact Words

<table>
<thead>
<tr>
<th>abase</th>
<th>advance</th>
<th>appear</th>
<th>avoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>abate</td>
<td>advertise</td>
<td>appease</td>
<td>awake</td>
</tr>
<tr>
<td>abort</td>
<td>aerate</td>
<td>apply</td>
<td>award</td>
</tr>
<tr>
<td>abrade</td>
<td>afford</td>
<td>argue</td>
<td>back</td>
</tr>
<tr>
<td>abridge</td>
<td>aggravate</td>
<td>arouse</td>
<td>bail</td>
</tr>
<tr>
<td>absorb</td>
<td>agree</td>
<td>arrange</td>
<td>bank</td>
</tr>
<tr>
<td>abstract</td>
<td>aid</td>
<td>arrest</td>
<td>bar</td>
</tr>
<tr>
<td>accelerate</td>
<td>aim</td>
<td>arrive</td>
<td>barbarize</td>
</tr>
<tr>
<td>accent</td>
<td>air</td>
<td>ask</td>
<td>bare</td>
</tr>
<tr>
<td>accept</td>
<td>allay</td>
<td>assemble</td>
<td>base</td>
</tr>
<tr>
<td>accredit</td>
<td>alleviate</td>
<td>assert</td>
<td>batter</td>
</tr>
<tr>
<td>achieve</td>
<td>alter</td>
<td>asseverate</td>
<td>beach</td>
</tr>
<tr>
<td>act</td>
<td>amend</td>
<td>assign</td>
<td>beam</td>
</tr>
<tr>
<td>add</td>
<td>amplify</td>
<td>assuage</td>
<td>bear</td>
</tr>
<tr>
<td>address</td>
<td>amuse</td>
<td>assure</td>
<td>become</td>
</tr>
<tr>
<td>adduce</td>
<td>animate</td>
<td>attach</td>
<td>befog</td>
</tr>
<tr>
<td>adjust</td>
<td>announce</td>
<td>attack</td>
<td>befuddle</td>
</tr>
<tr>
<td>administer</td>
<td>answer</td>
<td>attenuate</td>
<td>beget</td>
</tr>
<tr>
<td>admit</td>
<td>antedate</td>
<td>avert</td>
<td>begin</td>
</tr>
<tr>
<td>begrime</td>
<td>buy</td>
<td>compromise</td>
<td>damage</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>belch</td>
<td>bypass</td>
<td>conceal</td>
<td>damp</td>
</tr>
<tr>
<td>belie</td>
<td>canvass</td>
<td>concede</td>
<td>dance</td>
</tr>
<tr>
<td>bend</td>
<td>cap</td>
<td>conceive</td>
<td>dangle</td>
</tr>
<tr>
<td>benumb</td>
<td>capitalize</td>
<td>concuriate</td>
<td>darken</td>
</tr>
<tr>
<td>bequeath</td>
<td>carry</td>
<td>conclude</td>
<td>dary</td>
</tr>
<tr>
<td>bestow</td>
<td>cast</td>
<td>conduct</td>
<td>dash</td>
</tr>
<tr>
<td>betray</td>
<td>castigate</td>
<td>confess</td>
<td>deaden</td>
</tr>
<tr>
<td>better</td>
<td>castrate</td>
<td>confide</td>
<td>deal</td>
</tr>
<tr>
<td>bend</td>
<td>catch</td>
<td>confirm</td>
<td>debase</td>
</tr>
<tr>
<td>blackleg</td>
<td>chafe</td>
<td>confound</td>
<td>debauch</td>
</tr>
<tr>
<td>blanket</td>
<td>change</td>
<td>confuse</td>
<td>debunk</td>
</tr>
<tr>
<td>bleach</td>
<td>channel</td>
<td>congeal</td>
<td>decay</td>
</tr>
<tr>
<td>blemish</td>
<td>charge</td>
<td>connect</td>
<td>decide</td>
</tr>
<tr>
<td>blend</td>
<td>check</td>
<td>conserve</td>
<td>declare</td>
</tr>
<tr>
<td>blight</td>
<td>chill</td>
<td>consolidate</td>
<td>deepen</td>
</tr>
<tr>
<td>blister</td>
<td>chime</td>
<td>constitute</td>
<td>deface</td>
</tr>
<tr>
<td>block</td>
<td>chip</td>
<td>constrain</td>
<td>defeat</td>
</tr>
<tr>
<td>blockade</td>
<td>chock</td>
<td>constrict</td>
<td>defend</td>
</tr>
<tr>
<td>blow</td>
<td>choke</td>
<td>continue</td>
<td>deflate</td>
</tr>
<tr>
<td>blunder</td>
<td>choose</td>
<td>contort</td>
<td>deflect</td>
</tr>
<tr>
<td>blunt</td>
<td>churn</td>
<td>contract</td>
<td>deform</td>
</tr>
<tr>
<td>blur</td>
<td>cipher</td>
<td>control</td>
<td>defrost</td>
</tr>
<tr>
<td>blurt</td>
<td>circulate</td>
<td>convert</td>
<td>delay</td>
</tr>
<tr>
<td>bob</td>
<td>circumvent</td>
<td>convey</td>
<td>delegate</td>
</tr>
<tr>
<td>bog</td>
<td>claim</td>
<td>cook</td>
<td>deliver</td>
</tr>
<tr>
<td>boil</td>
<td>clash</td>
<td>cool</td>
<td>demise</td>
</tr>
<tr>
<td>bolster</td>
<td>clean</td>
<td>cordon</td>
<td>demonstrate</td>
</tr>
<tr>
<td>boost</td>
<td>cleanse</td>
<td>correct</td>
<td>dent</td>
</tr>
<tr>
<td>bowdlerize</td>
<td>clear</td>
<td>corrode</td>
<td>deny</td>
</tr>
<tr>
<td>bowl</td>
<td>climb</td>
<td>corrupt</td>
<td>deplete</td>
</tr>
<tr>
<td>brace</td>
<td>clinch</td>
<td>counter</td>
<td>depreciate</td>
</tr>
<tr>
<td>brand</td>
<td>clip</td>
<td>countersink</td>
<td>depress</td>
</tr>
<tr>
<td>brave</td>
<td>clog</td>
<td>cover</td>
<td>deprive</td>
</tr>
<tr>
<td>break</td>
<td>close</td>
<td>crack</td>
<td>depute</td>
</tr>
<tr>
<td>brief</td>
<td>clot</td>
<td>crank</td>
<td>derange</td>
</tr>
<tr>
<td>brighten</td>
<td>cloud</td>
<td>crash</td>
<td>describe</td>
</tr>
<tr>
<td>bring</td>
<td>cockle</td>
<td>craze</td>
<td>desecrate</td>
</tr>
<tr>
<td>broadcast</td>
<td>coin</td>
<td>create</td>
<td>design</td>
</tr>
<tr>
<td>bruise</td>
<td>collapse</td>
<td>cripple</td>
<td>designate</td>
</tr>
<tr>
<td>buckle</td>
<td>collect</td>
<td>crop</td>
<td>desolate</td>
</tr>
<tr>
<td>build</td>
<td>colour</td>
<td>cross</td>
<td>despoil</td>
</tr>
<tr>
<td>bull</td>
<td>comfort</td>
<td>crumble</td>
<td>destroy</td>
</tr>
<tr>
<td>burst</td>
<td>commission</td>
<td>crush</td>
<td>detail</td>
</tr>
<tr>
<td>bunch</td>
<td>commit</td>
<td>cry</td>
<td>detect</td>
</tr>
<tr>
<td>bundle</td>
<td>communicate</td>
<td>curb</td>
<td>deteriorate</td>
</tr>
<tr>
<td>bung</td>
<td>compare</td>
<td>curdle</td>
<td>determine</td>
</tr>
<tr>
<td>burlesque</td>
<td>complete</td>
<td>curtain</td>
<td>develop</td>
</tr>
<tr>
<td>burn</td>
<td>compound</td>
<td>cushion</td>
<td>die</td>
</tr>
<tr>
<td>burst</td>
<td>compress</td>
<td>cut</td>
<td>differentiate</td>
</tr>
</tbody>
</table>
diffuse earth exhale foil
dilute ease exhibit fold
dim eat exist follow
diminish educate expand force
direct effect expedite forgive
dirty elevate expose form
disable elicit express foster
disappear elude extend foul
discharge emancipate face found
discipline embitter extinguish frame
disclose embody extort fray
discolour emblazon fabricate free
disconnect emit fasten freeze
discontinue emphasize fasten frustrate
discover enable face furl
discuss encourage fade furnish
disfigure end fail furrow
disguise endorse fake fuse
dislocate enforce falsify gain
dislodge engage familiarize gallop
disconnect emphasize fasten garble
disrupt emit father gash
disappear enliven feature gerrymander
discharge emancipate father generate
disclose embitter extinguish gash
disorder emit father generate
discharge enliven feature gerrymander
discontinue emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
discontinue emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
disconnect emphasize fasten frustrate
disappear enliven feature gerrymander
disconnect emit fasten freeze
harm
harmonize
harry
hasten
hatch
head
heal
hear
heat
heighten
help
hide
hit
hoard
hoist
hold
hope
hound
hurt
identify
illuminate
imagine
impair
impart
impeach
impede
imperil
implant
improve
inaugurate
increase
indent
indenture
indicate
induce
induct
infest
infiltrate
infir
inflame
inflate
inflame
infect
influence
inform
infuse
initial
initiate
injure
insert
inspire
show
shrink
shut
sift
sign
signal
signalize
signify
simmer
sing
singe
sink
sit
site
situate
skirt
slacken
slake
slash
sleep
slice
slip
slow
smear
smile
smudge
snag
snap
snarl
snuff
sober
soften
soil
solace
solidify
soothe
sort
sound
sour
sow
spare
spark
speak
speck
speed
spill
spin
splinter
split
splodge
spoil
sponsor
spout
sprain
spray
spread
spring
square
squash
squeeze
stack
staff
stain
stalemate
stall
stamp
stand
star
starch
start
staunch
stay
steady
steer
stem
step
stick
stiffen
still
stir
stoke
stop
store
straighten
strain
strand
strengthen
stress
stretch
strike
strip
strum
study
stuff
stultify
stunt
subdue
subscribe
subvert
succeed
suffer
suggest
suit
summarize
supplement
supply
support
suppose
suppress
surface
surrender
survive
suspend
sustain
sweep
sweeten
swell
swing
swish
taint
tarnish
task
thank
thaw
thin
thrill
throw
thrust
thump
thwart
tidy
tighten
toll
tootle
topple
torment	
torture
total
touch
toughen	
tousle
tow
train
transfer
transplant
trap
treat
trigger
trim
truss
try
tumble
turn
twang
twiddle
twirl
twist
unblock
unburden
unclog
undo
unfasten
unfix
unfold
unhinge
unhitch
unite
unloose
unravel
unsaddle
unseat
unsex
unstop
untangle
untwist
uphold
upset
urge
use
validate
vandalize
veer
veil
ventilate
vocalize
voice
vote
vulgarize
waft
waggle
wake
<table>
<thead>
<tr>
<th>walk</th>
<th>wear</th>
<th>wilt</th>
<th>work</th>
</tr>
</thead>
<tbody>
<tr>
<td>wangle</td>
<td>weave</td>
<td>win</td>
<td>worry</td>
</tr>
<tr>
<td>warm</td>
<td>weep</td>
<td>wind</td>
<td>wreak</td>
</tr>
<tr>
<td>warn</td>
<td>weld</td>
<td>wing</td>
<td>wreck</td>
</tr>
<tr>
<td>warp</td>
<td>whet</td>
<td>wipe</td>
<td>wrest</td>
</tr>
<tr>
<td>warrant</td>
<td>whirl</td>
<td>wire</td>
<td>wring</td>
</tr>
<tr>
<td>wash</td>
<td>whitewash</td>
<td>wish</td>
<td>wrinkle</td>
</tr>
<tr>
<td>watch</td>
<td>widen</td>
<td>withdraw</td>
<td>write</td>
</tr>
<tr>
<td>weaken</td>
<td>wield</td>
<td>wither</td>
<td>yield</td>
</tr>
<tr>
<td>wean</td>
<td>wiggle</td>
<td>withhold</td>
<td></td>
</tr>
</tbody>
</table>
1. A method of finding facts (610) within electronic resources (116), comprising:
scanning (502) an electronic resource (116) to discover factual descriptions
(402) of sentences that comprise words matching words of a fact-word table;
examining (506) the discovered factual descriptions (402) to identify the
linguistic constituents of the factual descriptions; and
determining (510) whether to present a factual description (402) as a fact (610)
based on the identified linguistic constituents.

2. The method of claim 1, wherein determining whether to present a factual
description as fact based on the identified linguistic constituent comprises:
applying excluding rules in relation to the linguistic constituents of the factual
descriptions to eliminate certain factual descriptions from consideration;
scoeming the factual descriptions;
comparing the score of each factual description remaining for consideration to a
threshold; and
for each factual description having a score that exceeds the threshold, presenting
at least a portion of the sentence containing the factual description as a fact.

3. The method of claim 2, further comprising tagging words of the factual
descriptions with their parts of speech.

4. The method of claim 3, wherein tagging words of the factual descriptions with
their parts of speech comprises applying a noun tag when a word may be either a verb
or a noun.

5. The method of claim 4, wherein applying the excluding rules comprises
applying a first set of rules for syntactic phrases that have a role of subjects and
applying a second set of rules for syntactic phrases that have a role of objects.
6. The method of claim 5, wherein applying the first set of rules comprises excluding noun phrases having an opinion or biased modifier of subjects or objects.

7. The method of claim 5, wherein applying the second set of rules comprises excluding subject noun phrases which non-proper name definite descriptions, excluding noun phrases which contain pronouns, and excluding subject noun phrases which do not appear at the beginning of text.

8. The method of claim 5, further comprising applying a third set of rules without regard to the role of the noun phrase.

9. The method of claim 8, wherein applying the third set of rules comprises excluding factual descriptions where the punctuation of the sentence is a question mark, and excluding sentences with phrases that include a stop word.

10. The method of claim 2, wherein scoring the factual descriptions comprises scoring only those factual descriptions remaining for consideration either after or during application of the excluding rules.

11. A computer readable medium (112) containing instructions (120) that perform acts comprising:

   receiving a search term (206);
   parsing (502) a plurality of electronic documents (116) to discover factual descriptions (402) of sentences that comprise words matching words of a fact-word table;
   examining (506) the discovered factual descriptions (402) to identify the linguistic constituents of the factual descriptions (402); and
   determining (510) whether to present a factual description (402) as a fact (610) relevant to the search term based on the identified linguistic constituent.

12. The computer readable medium of claim 11, wherein the acts further comprise obtaining the plurality of documents by searching an collection of electronic documents to find those documents containing the search term, wherein the collection is searched
13. The computer readable medium of claim 11, wherein the acts farther comprise obtaining the electronic documents and presenting factual descriptions prior to receiving the search term and searching the electronic documents and factual descriptions to find those electronic documents and corresponding factual descriptions that are relevant to the search term.

14. The computer readable medium of claim 11, wherein determining whether to present a factual description as a fact relevant to the search term based on the identified linguistic constituent comprises:
   - applying excluding rules in relation to the linguistic constituents of the factual descriptions to eliminate a portion of the factual descriptions from consideration;
   - scoring the factual descriptions;
   - comparing the score of each factual description remaining for consideration to a threshold; and
   - for each factual description that is taken from an electronic document that contains the search term and that has a score that exceeds the threshold, presenting at least a portion of the sentence containing the factual description as a fact relevant to the search term.

15. The computer readable medium of claim 14, wherein scoring the factual descriptions comprises scoring only those factual descriptions remaining for consideration after applying the excluding rules.

16. A computer system (100), comprising:
   - storage (112) containing a plurality of electronic resources (116) that comprise textual information;
   - a processor (102) that receives a request (206) to present facts that are related to the search term from a set of electronic documents, wherein the processor (102) parses (502) the plurality of electronic documents (116) to discover factual descriptions (402) of sentences that comprise words matching words of a fact-word table, examines (506)
the discovered factual descriptions to identify the linguistic constituents of the factual
descriptions (402), determines (510) whether to present a factual description (402) as a
fact (610) based on the identified linguistic constituent, and presents (514) at least a
portion of sentences that contain the factual descriptions (402) that are determined to be
presented as a fact (610) and that are related to the search term (206).

17. The computer system of claim 16, further comprising a display device and
wherein the processor presents at least the portion of the sentences by displaying at
least the portions of the sentences on the display device.

18. The computer system of claim 16, further comprising a network interface and
wherein the processor presents at least the portion of the sentences by outputting those
portions to another computer via the network interface.

19. The computer system of claim 16, further comprising a network interface and
wherein the storage is accessible by the processor via the network interface.

20. The computer system of claim 16, wherein the processor determines whether to
present a factual description as fact by:
   applying excluding rules in relation to the linguistic constituents of the factual
descriptions to eliminate a portion of the factual descriptions from consideration;
   scoring the factual descriptions;
   comparing the score of each factual description remaining for consideration to a
threshold; and
   for each factual description that contains the search term and that has a score
that exceeds the threshold, presenting at least the portion of the sentence containing the
factual description as a fact relevant to the search term.
FIG. 2

1. Obtain collection of electronic documents
2. Perform fact extraction on the electronic documents
3. Receive search term(s)
4. Find relevant electronic documents
5. Display list of relevant electronic documents and relevant facts

FIG. 3

1. Receive search term(s)
2. Find relevant electronic documents
3. Perform fact extraction limited to relevant facts from relevant electronic documents found
4. Display list of relevant electronic documents and relevant facts
FIG. 4

RECOGNIZE FACTUAL DESCRIPTIONS

PERFORM FACT EXTRACTION ON THOSE FACTUAL DESCRIPTIONS

FIG. 5

SCAN DOCUMENT FOR FACTUAL DESCRIPTIONS MATCHING THOSE OF FACT WORD TABLE

TAG WORDS OF FACTUAL DESCRIPTIONS WITH PART OF SPEECH

IDENTIFY SYNTACTIC PHRASES IN THE NEIGHBORHOOD OF EACH FACTUAL DESCRIPTION AND IDENTIFY THE ROLE OF THE SYNTACTIC PHRASES

APPLY SUBJECT, OBJECT, AND SENTENCE BASED EXCLUSION RULES TO THE SYNTACTIC PHRASES FOR THE FACTUAL DESCRIPTIONS

FOR REMAINING SYNTACTIC PHRASES FOR THE FACTUAL DESCRIPTIONS, APPLY RULES TO SCORE EACH FACTUAL DESCRIPTION

SCORE EXCEED THRESHOLD?

PRESENT SENTENCE / PARAGRAPH AS FACT

DISCARD FACTUAL DESCRIPTION

NO

YES
Welcome to the Official John Doe Web Site
Copyright 2006 John Doe, Inc. All rights reserved
www.john doe.com  Cached pages

John Doe | Gallery
Gallery 1
www.john doe.com/gallery/gallery.asp  Cached page
+ Show me more results from this site

CMT.com: John Doe: Artist Main
CMT.com presents complete John Doe information including Music Videos, Photo Galleries, News, Radio, TV Appearance, Tour Dates, Discography, Message Board, Awards and more.
www.cmt.com/artist/john_doe/artist.htmlCached page

John Doe - Wikipedia, the free encyclopedia
John Doe (born July 4, 1960) is a successful American country music singer. He now resides in Memphis, Tennessee off of John Doe Highway. Doe was born in Nashville, Tennessee to Jim and Joan.....
en.wikipedia.org/wiki/John_Doe Cached page

John Doe
Black 100% cotton shirt featuring the John Doe winged logo on the front and the list of cities he played during his 2004 tour.
City: 1 2 3 4 5 6 7 & Size: L XL
www.fanfire.com/artist/John%2BDoe&sourceCode=TTRWEB Cached page

John Doe Fan Tribute
Doe released a holiday album in 1992, A John Doe Christmas: Loving Time Of The Year. His autobiography, Walking and Talking, was released in conjunction with the album of the.....
artwork.net/lace/john_doe.htm Cached page

John Doe
DAYTONA BEACH, FL - The owner of the Legendary Florida Campground announced today that Country Music Legend John Doe will play March 5 at Kickstand City in the infield of Daytona ...

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/21(2006.01), G06F 17/27(2006.01)

According to International Patent Classification (IPC) or to both national classification and IPC

B. MINIMUM DOCUMENTATION SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 8 G06F

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

Korean Utility models and applications for Utility models since 1975
Japanese Utility models and applications for Utility models since 1975

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PAI, FPD, USPAT, eKIPASS, IEEE, YAHOO, GOOGLE, Keyword "fact extraction, Information extraction, scoring, exception rule"

C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>US 2005-0108630 A1 (MARK D WASSON et al) 19 May 2005 (See Abstract, paragraphs [0134]-[0143], [0208]-[0517])</td>
<td>1-20</td>
</tr>
<tr>
<td>A</td>
<td>MARIUS PASCA et al Organizing and Searching the World Wide Web of Facts Step One the One-Million Fact Extraction Challenge Proceedings of the Twenty-First national Conference on Artificial Intelligence, 16-20 July 2006, Published by The AAAI Press Pages 1400-1405 (See Abstract, right column of page 1400-left column of page 1402)</td>
<td>1-20</td>
</tr>
<tr>
<td>A</td>
<td>MARIUS PASCA et al Names and Similarities on the Web Fact Extraction in the Fast Lane Proceeding of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, July 17-21 2006, pages 809-816 (See Figure 1, Section 2 Contributions, Section 2 Generalization via Word Similarities)</td>
<td>1-20</td>
</tr>
<tr>
<td>A</td>
<td>US 6741986 B2 (RAYMOND J CHO et al) 25 May 2004 (See Abstract, Figs 3-6,column 2 line 28-column 3 line 35, column 7 lin 37-column 22 line 11)</td>
<td>1-20</td>
</tr>
</tbody>
</table>

Date of the actual completion of the international search

03 DECEMBER 2007 (03 12 2007)

Date of mailing of the international search report

03 DECEMBER 2007 (03.12.2007)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
920 Dunsan-dong, Seo-gu, Daejeon 302-701, Republic of Korea
Facsimile No 82-42-472-7140

Authorized officer

SUN Dong Guk
Telephone No 82-42-481-8248

Form PCT/ISA/210 (second sheet) (April 2007)
<table>
<thead>
<tr>
<th>Patent document cited in search report</th>
<th>Publication date</th>
<th>Patent family member(s)</th>
<th>Publication date</th>
</tr>
</thead>
<tbody>
<tr>
<td>US20050108630A1</td>
<td>19.05.2005</td>
<td>AU2004294094A1</td>
<td>09.06.2005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA2546896AA</td>
<td>09.06.2005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP01695170A2</td>
<td>30.08.2006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO2005052727A2</td>
<td>09.06.2005</td>
</tr>
<tr>
<td>US6741986B2</td>
<td>25.05.2004</td>
<td>AU2002340393BB</td>
<td>18.01.2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU2006201478AA</td>
<td>27.04.2006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA2465592AA</td>
<td>22.05.2003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP01454264A1</td>
<td>08.09.2004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP1454264A1</td>
<td>08.09.2004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US06741986</td>
<td>25.05.2004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US20030074516A1</td>
<td>17.04.2003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US2004220969A1</td>
<td>04.11.2004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US2004236740A1</td>
<td>25.11.2004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US2005055347A9</td>
<td>10.03.2005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US6772160BB</td>
<td>03.08.2004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO2003042872A1</td>
<td>22.05.2003</td>
</tr>
</tbody>
</table>