Patent Semantics: Analysis, Search, and Visualization of Large Text Corpora

by

Christopher G. Lucas

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degrees of

Bachelor of Science in Computer Science and Engineering

and Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

August 20, 2004

Author____
Department of Electrical Engineering and Computer Science
August 20, 2004

Certified by____
Deb K. Roy
Thesis Supervisor

Accepted by____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Patent Semantics: Analysis, Search, and Visualization
of Large Text Corpora

by
Christopher G. Lucas

Submitted to the
Department of Electrical Engineering and Computer Science

August 20, 2004

In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer [Electrical] Science and Engineering
and Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

Patent Semantics is system for processing text documents by extracting features capturing their semantic content, and searching, clustering, and relating them by those same features. It is set apart from existing methodologies by combining a visualization scheme that integrates retrieval and clustering, providing a variety of ways to find and relate documents depending on their goals. In addition, the system provides an explanatory mechanism that makes the retrieval an understandable process rather than a black box. The domain in which the system currently works is biochemistry and molecular biology patents but it is not intrinsically constrained to any document set.

Thesis Supervisor: Deb Roy

Title: Associate Professor, MIT Media Laboratory

1. INTRODUCTION

When most people think of search technology, the first thing that comes to mind is web search, as exemplified by Google's search engine and its competitors. These technologies have a few common characteristics, including presenting their results as sorted lists, searching for documents in one way, and keeping the algorithms behind their search secret. For many web searches, these are useful properties: people looking for web pages often wants a single typical page about something and would rather not be bothered with any parameters past a couple of words, information describing the search, or even having to pick a result (as per Google's "I'm feeling lucky" feature).

Not are searches are the same, however. There are times when users want to retrieve a lot of results and see where the good ones end and the bad ones begin, or to see relationships between results, making an ordered list of results inappropriate. There are times when a user's criteria for good results do not fit into a mold used for every other search, making a single black-box search algorithm inappropriate. Finally, a user may be searching for something that is elusive because he or she finds it difficult to capture the semantics of the data in mind in words, at least without also getting so much chaff that the results are useless. In this case, a good way to separate the good from the bad is to learn what the nature of the bad is and account for it, suggesting that it would be useful for users to understand why search returned what it did.

The Patent Semantics system tries to address these deficiencies by letting users search in a way that is visual, flexible and self-explaining.

1.1 Project History

Patent Semantics is the evolution of a project that started with the goal of building complete graphical descriptions of the operations described in biochemical synthesis patents. The project has been driven by three objectives:

- To build a patent search system that outperforms the standard tool for patent search, the US Patent and Trademark Office's web-based boolean search.
- To make the basis for its retrieval of a given patent clear to the user, permitting informed query refinement. This is in contrast to the take-it-or-leave-it results returned by most search systems today, which require the user to guess appropriate synonyms and context-providing terms when an initial query fails.
- To let a machine decouple the syntactic and lexical details from the semantics of procedure-describing documents.

The first pass at building a system involved using information from Framenet II roles and frames (Baker, Fillmore, Lowe, 1998) to resolve ambiguous anaphora, elliptical references, and rule out impossible interpretations of input texts. For instance, given temporally ordered consecutive sentences A,B,C,D,E, the system would determine that a substance produced in B and consumed in D cannot have any role in A or E. A large number of potentially-valid assignments of objects to semantic roles were enumerated and scored by the appropriateness of their basic type to the role, according to templates established a priori, to build graphical representations of procedures from text. Had it been successful, the result of such an analysis would be a graph that captured dependencies between various substances and procedures used in the synthesis protocol.

After some exploratory work it became apparent that the approach was impractical using currently available semantically or syntactically annotated corpora. Even if the role-assigned system had functioned perfectly, there were important distinctions between the semantic correlates of lexical forms that were not captured by Framenet II or any other useable data set. For example, correctly interpreting a synthesis patent may require a chemist to be aware of properties of RNA polymerases that are not articulated in any machine-readable form.

The current form of the project is faithful to the spirit of the original goals, attempting to capture and articulate the semantics of documents and the relationships between them, while not requiring that its underlying tools perform flawlessly. It was conceived as a step towards an architecture that represents and relates documents in ways compatible with human intuition as well as a useful tool for information retrieval and organization.

1.2 Prior Art

A discussion of selected relevant previous work in the domains of information retrieval and information visualization follows.

The US Patent and Trademark Office has an online patent search tool that is a common starting place for people who are interested in finding patents. Its most advanced search comprises a text field in which people can compose boolean queries over the contents of various fields of patents. The boolean query approach fails entirely to address the problems of synonymy and polysemy; when a word has a synonym, the search can never retrieve documents that contain the synonym and lack the exact query term. Simultaneously, if the query term has an alternate meaning, a query will return every document with that meaning. These issues can sometimes be addressed by composing queries that explicitly state synonyms and rule out different meanings with additional terms that refine the context of the query terms, but there is no recourse if the user is unaware of an important synonym or secondary term that distinguishes contexts.

Leah Larkey (1999) built a system for search and classification of patents that used a k-nearest-neighbor classifier to label patents with US classifications and a query expansion tool using noun phrase co-association probabilities. Larkey's classifier classifies new patents by using their text as a query to a black-box search engine, labeling it with the most common class in an N-best results retrieved. Larkey's query expansion system locates noun phrases by segmenting sentences on predefined punctuation and word boundaries. It assigns cooccurrence scores for every phrase pair {a,b} using a function that is monotonically increasing with P(a,b)/(P(a)P(b)) and uses the phrases that score highest against query phrases to suggest query expansions. Larkey's system does not classify or group patents more precisely than US classifications permit, and her system does not account for cases in which different synonyms for a term are not used interchangeably, as is often the case when two distinct communities develop technologies that overlap in form or function.

Omniviz is a commercial tool for mining text documents, and it is used by companies to understand patterns in patent data. It has various tools for visualizing patent data, including a scatter plot that groups patents using k-means clustering, and a simple summarization tool. It has tools to look at correlations between features, to view query results in a handful of different ways, such as in a polar coordinates, and to summarize documents using a handful of words. It is chiefly a visualization tool, and does not purport to introduce new approaches to retrieval (OmniViz, 2004).

IPVision is a company that creates visualizations of the relationships between patents. Its main product is a diagram of the citation links between patents and patent portfolios of various organizations. Unlike this project, IPVision's approach does not address the content of patents, instead its technology concentrates almost exclusively laying out and grouping patents by their owners and citations. (IPVision, 2004).

Personal correspondence with practicing patent attorneys and intellectual property managers indicated that the tools available for finding and organizing patents are not satisfactory because the deliberately vague and obfuscated language of patents makes keyword-based searching extremely time-consuming, and because the more sophisticated tools do not provide high enough recall rates are too opaque in their operation to trust.

1.3 Organization

The remainder of this thesis is organized into three sections. A high-level overview of the system follows is given first. Second, the system is described in detail. Finally, the system's performance and future directions are discussed.

2. CONCEPTUAL OVERVIEW

When a user executes a query on a search engine, s/he has some goal in mind. In the case of most web queries, a user is looking for a small number of pages that are all about the subject of his query. Google's search engine is tuned to such goals, ranking pages based on their authority as inferred from hyperlinks and the extent to which query terms appear is headers and other emphasized text (Brin and Page, 1998). In the case of patent search, a user's goals are more diverse. One user might be a patent attorney with an invention description in hand, looking for claims in existing patents that might make writing a patent for the invention a waste of time. In this case, the user wants a system that provides good recall without forcing him to sift through thousands of weak matches, and doesn't care about how the results relate to one another or where the patents came from. A different user might be someone forming a corporate research plan, who wants to see who potential competitors are in different area by their patent portfolios, painted in broad strokes. This person wants something that returns an accurate gist that can be scanned an understood in seconds, but she does not care if the system misses a patent here and there. A third user might be a manager in charge of finding what technology licensing is necessary to see a product, and would want to know who has what patents for components of the product. These are only a few of the ways in which users might be interested in looking for and looking at patents, suggesting that in the domain of patents a 'one-search-fits-all' philosophy is inappropriate.

The Patent Semantics system responds to this problem by providing users with a diversity of lenses through which they can force the system look at patents, and helping them understand each lens in the context of their queries so as to use the system's versatility to the maximum effect. It integrates three distinct kinds of features that are quite different from one another and are useful in different contexts. The term 'lens' is used because different feature sets modifies how a user views documents in some consistent way like a lens modifies of someone views physical objects. One feature set will clarify some things while suppressing others that may be irrelevant to the user's goals – much like a choice of lenses influences what can be seen, based on depth of focus, extent of zoom, and selective distortion.

The first feature set is a weighted set of word frequencies in documents. Words provide an intuitive basis for queries; from the USPTO's boolean queries to Google, the search engines that see the most use return results by matching words in a query to words in documents. In cases where a user can think of words that define a concept precisely and completely, no other features are necessary – at least for retrieval. This lens allows users to bring out documents containing lexically unambiguous sets of terms in sharp contrast, while blocking out documents that do not, but has blurry and blind spots in the case of polysemous and synonymous words, respectively.

The second feature set comes from a well-established technique in information retrieval, called Latent Semantic Analysis (LSA). It comprises an orthonormal set of components produced by the singular value decomposition of a weighted term-by-document matrix. Its advantages over words come from the fact that its components contain groups of words that relate to one another by way of occurring together, hence the words 'latent semantics' in its description. For example, a search for "heart disease" using just the LSA distance measures in the Patent Semantics system gives a patent that is clearly relevant to heart disease but never uses the word heart, so it's never captured in a word-based search. LSA-based features also provides for fewer misses when relating documents to one another, potentially giving better results when clustering documents by their themes. An LSA-based representation affords a global view of data that is less prone to blind spots than a lexical one, but it does not provide as clear a picture if there are unambiguous terms with which to work, and the mapping from the underlying for of the data to the view not as immediately obvious.

The third feature set is composed of unique concept identifiers that are linked to noun and verb phrases in the patents. These concept identifiers come from the National Library of Medicine's Unified Medical Language System Metathesaurus and Semantic Net (Nelson, Powell, Humphreys, 2002) and were created by several communities of experts in biology and medicine. Like the LSA feature set, it permits generalization of queries, using explicit relationships between concepts, but it has the advantage of being more selective. For instance, if a user gives a query "DNA polymerase" LSA finds strong linkages between that and RNA polymerase, as well as a number of other terms that are linked to but not synonymous with DNA polymerase, while UMLS contains clear distinctions between DNA polymerase and RNA polymerase. It has its own disadvantages, which fall into two categories: appropriateness and coverage. The problems of appropriateness stem from the fact that any conceptual organization scheme, or ontology, is made by people with particular goals in mind, and those goals do not necessary overlap with all of the system's users. For example, a user might execute a query with the hope of finding patents related to a drug by its chemical structure, while the ontology may give results that are similar based on the classes of diseases it treats. Lack of coverage is also a result of a mismatch between the goals of the builders and users of the ontology. For example, the medically-oriented UMLS covers stents as examples of devices, but does not cover microfluidic wells that are used for experimentation by pharmaceutical companies. Some ontologies, like Wordnet (Fellbaum, 1998) have very broad coverage that includes both ‘mitochondria’ and ‘justice’, but at the expense of making potentially important distinctions between very specific things like classes of enzymes. This lens allows users to view certain aspects of patents in a very clearly, with well-articulated context, while sacrificing the parts it does not cover wholesale.

Feature sets can be used for document retrieval by way of various distance measures defined on them. The same distance measures can be coupled to a dimensionality reduction tool to cluster patents, allowing users to get a sense of how they form distinct classes. Seeing such classes allows users to understand the general directions in which work in a field is going, and to find batches of patents that are related to a topic of interest.

After clustering or retrieval, users can obtain explanations from the system for documents’ relationships with one another and with queries, for any of the features sets. A user might through this functionality discover that unwanted documents are being returned by the UMLS because of alternate meanings for a word in her query phrase. She might simultaneously discover a more discriminating synonym, allowing her to get better retrieval performance.

3. DETAILED SYSTEM DESCRIPTION

This section is a detailed description of the parts of the system summarized above. It includes the structure, creation, and hardware behind of the data store, a listing of all of the parts of the back end, and an illustrated walkthrough of a query on the front-end.

3.1 Hardware

The database server machine is an Intel Pentium 4 running Red Hat Enterprise 3.0, with 4GB of physical memory. The patent data are housed in a MySQL database on two 250GB SATA drives in a RAID 0 configuration.

3.2 Data

The main database contains biochemistry-related (US class 435) patent data for the years 2002 and 2003, stored in a patents table, a classification table, and a phrase feature table. It was built by parsing XML files purchased from the US Patent and
Trademark Office, and covers 9875 patents.

An XML parser reads raw data that had a one-to-one correspondence with patent numbers into a one table, storing the title, abstract, claims, assignment date, and assignee for a patent, indexed by patent number.

The parser then reads raw data with many-to-one mappings onto patent numbers into distinct tables. These tables include patent numbers for the other patents that a particular patent cites, and the US classifications of each patent, of which there are sometimes many.

The final table the parser builds is a collection of derived syntactic and semantic features. After reading the claims and abstracts of each patent, the parser tokenizes the text using a perl script that splits contractions, separates punctuation from words, e.g., “Don’t talk, smoke, or chew gum” becomes “Don ‘t talk , smoke , or chew gum”, , tags each word and symbol by its part of speech such as singular noun, gerund, or pronoun, using Eric Brill’s tagger (Brill, 1992), and uses the Collins parser (Collins, 1999) to label and extract syntactic elements. The output of the part of speech tagger on a sample sentence looks like this:

```
a/DT plurality/NN of/IN nanoparticles/NNS bonded/VBN to/TO the/DT bridging/VBG DNA/NNP ./.
```

Slashes are delimiters, and almost all of the DT, NN, and other tags are in the Penn TreeBank tag set that the Collins parser uses. Those that are not are converted to their closest matches using regular expressions.

The output of the Collins parser for the Brill output above is

```
(TOP (S (NP-A (NPB (DT a) (NN plurality)) (PP (IN of) (NPB (NNS nanoparticles)))) (VP (VBN bonded) (PP (TO
to) (NPB (DT the) (VBG bridging) (NN DNA) (. .))))))
```

The elements extracted from parses include verbs and noun phrases, and are labeled in the table by where they appear, by their phrase type, and by concepts in the UMLS ontology that they describe. The concepts are found by looking up case-insensitive matches in the surface forms stored in the UMLS Metathesaurus database. Some verbs and noun phrases and with linked concept identifiers:

```
| NOUN | risk               | C0035647 |
| NOUN | multiple sclerosis | C0026769 |
| NOUN | nucleic acid       | C0028606 |
| NOUN | 6                  | C0205452 |
| NOUN | oligonucleotide    | C0028953 |
| NOUN | allele             | C0002085 |
| NOUN | 1                  | C0205447 |
| NOUN | respect            | C0679133 |
| VERB | is                 | C0441913 |
| VERB | using              | C0439224 |
| NOUN | type 1 diabetes    | C0011854 |
| NOUN | c                  | C0439128 |
| NOUN | genotype           | C0017431 |
| NOUN | a                  | C0439126 |
| NOUN | nucleotide         | C0028630 |
| VERB | based              | C0178499 |
```

The decision to extract noun phrases and verbs rather than other syntactic elements has two justifications:

- The extracted elements should have corresponding strings in the UMLS Metathesaurus. Noun phrases have better coverage than nouns, and verbs have better coverage than verb phrases.
- The extracted elements should cover a large proportion of the patent text while having minimal overlap. If verb phrases had been used, then the verb elements would subsume a large proportion of the noun phrases, while noun phrases rarely subsume verbs.

Noun phrases at all parse depths were extracted, rather than simply top-level noun phrases. For example, the noun phrase: “outer surface of layer of ceramic material” will also enter the database as “layer of ceramic material” and “ceramic material”. This approach gives better UMLS coverage than would using only top-level phrases. In total, 1612502 phrases were extracted, 449737 of which were attached to concepts, including 8509 distinct concept identifiers.

For a concise summary of the tables in the tables, see appendix A. For more details on the natural language processing involved in extracting phrases, see appendix B. For more details on the tables contained in the UMLS database, see appendix C. For references to the source files for parsing, database management, and other data storage functionality, see appendix D.

3.3 Tools for Retrieval and Analysis

The text of patents and the phrase-based data in the database are used to compute distances between patents and distances between queries and queries and patents, create summaries of how these distances are computed for the user, and visualize how these patents relate to one another by dimensionality reduction and clustering.

A first step in all of this is the creation of feature sets that give good performance in retrieving that relating patents, and in creating explanations of relatedness that make sense to the user. The patent semantics system uses three feature sets that satisfy these needs in different ways: weighted frequencies of individual words, latent semantic components obtained via Latent Semantic Analysis techniques, and linkages from patent text to concepts in the UMLS Metathesaurus. These features and the distance functions used for them will be described in turn, followed by a description of the algorithms for dimensionality reduction, clustering, and pairwise relatedness explanation.

3.3.1 Lexical Distance

The weighted word frequency is the simplest of the three distance measures, provides a baseline query system and allows fast query-based filtering. Documents are reduced to vectors describing the logs of their unordered word counts, weighted by *1-entropy(term)* This is the same weighting used to build a term-by-document matrix for the LSA measure and is discussed in (Dumais, 1991). It is described by the equation

$$1+\sum_{j}\frac{p_{ij}\log p_{ij}}{\log ndocs}$$

where p_{ij} is equal to tf_{ij}/gf_{ij}, tf_{ij} defined as the frequency of term i in document j, and gf_{ij} is the global frequency of term j, and *ndocs* is the total number of documents in the corpus.

Using this weighting, terms that appear in almost every document are heavily discounted, while terms that appear in only a handful of documents are weighted very heavily. In the extreme case, a word that appears only once would be given a weight equal to 1+[log(1)/log(*ndocs*)], or one. As examples of the actual weights computed, the word ‘a’ has a weight of 0.06, while the word ‘aberrant’ has a weight of 0.88. This entropy-based measure was chosen because it is both an intuitively reasonable to weight terms, like the well-known tfidf measure but with more principled underpinnings in information theory, and demonstrated to be a superior weighting scheme for LSA in (Dumais, 1991).

The similarity of these weighted word vectors is computed using their cosine, equal to their inner product divided by the product of their norms. It returns one for identical documents and zero for documents with no shared terms.

$$S(V_1,V_2) = \frac{\langle V_1,V_2 \rangle}{|V_1||V_2|}$$

3.3.2 LSA

Latent Semantic Analysis is an approach to information retrieval built around a decomposition of a documents-by-terms matrix into two orthonormal matrices and a diagonal matrix of singular values by way of the Singular Value Decomposition algorithm. It describes documents as sets of components that are weighted subsets of the vocabulary. One view of these components is as semantically coherent and significant themes across the body of documents. LSA permits mapping novel documents into the component space by way of projection, so it is not necessary to compute the SVD for every document to be retrieved

The documents-by-terms matrix is built using the log-entropy weighting scheme for words described above, using 7000 words and 7000 documents. A square matrix is not required by the algorithm – the distinct words need only be equal to or greater than the number of documents- these numbers were chosen because the tools used to calculate the SVD – CERN's COLT library for java, and standard matrix templates for C - were incapable of handling larger matrices on the hardware available. Computing the SVD on this matrix required roughly 1.5GB of memory and over 25 hours of machine time. The resulting three matrices were stored as tab-delimited text files, and then serialized as a java object for rapid reading. The 7000 documents were chosen arbitrarily using a ‘limit 7000’ modifier on a SQL query to the database, and the words were those high the highest tfidf scores – global term frequencies divided by the number of documents in which the terms appeared. Attempts were made to work with larger matrices, ideally one covering the total vocabulary of 35,000 for roughly 10,000 patents, by using efficient matrix libraries such as Intel’s Math Kernel Library. This was not possible in the time available.

The semantic coherence of the components computed was informally tested, producing mixed results. Some components were characterized by sets of clearly related words: the second most significant component had three-letter acronyms for amino acids such as “leu” and “ser” for its fifteen most heavily weighted word constituents. Another was best described by the words “plant”, “seed”, “corn”, “inbred”, “plants”, “soybean”, “hybrid”, and “crossing”. Unfortunately, the coherence of components was not uniform; the majority was characterized by words of irregular similarity. For example, the most heavily weighted words in the most significant component were “said”, “acid”, “sequence”, “c”, “nucleic”, and “id”. It is likely that failures such as these are a symptom of sparse data, caused by coincidentally high co-occurrences of words. Using a larger data set such that distinct themes of patents wash out lexical noise should reduce or eliminate the problem.

Only a subset of the components for a document are used, usually the 100-300 with the largest singular values. If all of the components were used, the LSA distances between documents are equal to the lexical distances, at least for documents used in the original SVD computation. The intuition behind this reduced dimensionality is that the real meanings of documents are captured by the principal components and the remainder is lexical chaff. In practice, it allows documents to be retrieved that contain words related to a query word that are not the query word itself. For instance, searching for “corn” might pick up documents about soybeans by way of the plant-related component described above. The LSA-based distance measure uses a vector cosine between the components, but normalizes the result to a zero-to-one range by adding one and dividing by two; because of the presence of negative component weights in the SVD, a negative cosine is possible. After normalization, a similarity value of 0.5 suggests that documents are related at chance levels, while a value of zero suggests the documents’ contents are mutually exclusive in some non-random way.

3.3.3 UMLS

The UMLS Metathesaurus and Semantic Net, organizes a diverse collection of concepts that relate to medicine. These include but are not limited to: chemicals such as 1,2-Dipalmitoylphosphatidylcholine, disorders such as influenza, organisms such as E. coli, elements of living things from mitochondrial DNA to torsos, procedures performed by doctors and biologists such as blood coagulation tests, and concepts like humidity and normal. Unfortunately, it is an agglomeration of data assembled by different organizations, and it suffers from some structural inconsistencies. These cause problems in trying to get useful distances between terms for many ontological distance measures, including the Jiang-Conrath distance measure used in the Patent Semantics system. The Jiang-Conrath measure was chosen based on Budanitsky and Hearst’s (1991) comparison of several measures, where it was judged to be the best overall in terms of its consistency with human judgments.

JC distance takes advantage of corpus statistics in addition to the structure of the ontology being used, by calculating the distance between two concepts based on the probability that one of them is an example of their most specific shared parent in the corpus. Part of the intuition behind this approach is that a short path in an ontology need not imply similarity. For example, if an ontology has only two tiers – ‘thing’ and ‘example-of-a-thing’, it makes little sense to conclude from the lack of separation between ‘justice’ and ‘apple’ that the two are synonymous. The equation describing the JC distance between two concepts is

$$dist_{jc}(c_1 : c_2) = 2\log p(lso(c_1 : c_2)) - \log(p(c_1)) - \log(p(c_2))$$

where c_1 and c_2 are the concepts being related, and $lso(c_1 : c_2)$ is their most specific shared parent. This approach to measuring distance judges ‘justice’ and ‘apple’ as examples of ‘things’ to be different because the log-probability of the most-specific parent (thing) is zero, and the log-probability that any arbitrary thing is an apple or justice is a large negative number, leading to a high JC distance. The structural problems referred to above come into play when to unrelated terms are judged to be too close by any one of the many data sources in the UMLS Metathesaurus.

For example, if one data source happens to suffer from sparse coverage for a class of terms, e.g., fruits include only apples and oranges, or it contains an atypical small class, e.g., a class like ‘flies and has a name starting with the letter A’ then there will be an erroneously high similarity between some terms. This occurs most often in conjunction with problems of polysemy, where verbs or noun phrases have multiple meanings, one of which is very uncommon is the data set.

The Jiang-Conrath distance acts on pairs of concepts, so additional steps are necessary to turn it into a distance measure for documents that are linked to several concepts. First, mappings are found between all concepts in two documents. For the sake of simplicity and computational tractability, a greedy one-to-one mapping was chosen, choosing the lowest-distance pairing, removing those concepts from consideration, choosing the next-best pairing, until one document has no remaining concepts. At this point, there is a set of (concept1,concept2,distance) triples and a set of leftover concepts. There are various ways to relate documents using these. One of the simplest, and the one currently in use is to use the average similarity between concepts, ignoring those that remain in the document with more concepts after matching. A more complicated approach that is more sensitive to leftover concepts but was rejected because it fails more dramatically in cases of sparse coverage is to convert the mappings to a cosine-friendly representation. In this case the inner product of the documents is computed by using weighted matches. Identical concepts have a distance of zero, so every match weighted by 1/(1+distance). Suppose document A has concepts C0, C1, and C2, and document B has concepts C0, C5, C6 and C7, and that this matching produces the triples (C0,C0,0), (C1,C6,1), and (C2,C5,19), with C7 left over. Without introducing the addition complexity of log-entropy weighting as with terms, the 'pseudo-cosine' distance between these documents is (1/1 + 1/2 + 1/20)/(sqrt(3)*sqrt(4)), or (1.55/3.73). If all the concept matches had been perfect, the distance would have been 3/3.73, and if all has been as bad as the worst, which is to say the documents were almost totally unrelated, it would have been 0.15/3.73.

3.4 Dimension Reduction

Regardless of the distance measure used, patent documents fall into a space that is much higher than three dimensions. In an effort to see how patents fall into distinct groups and subgroups for the various feature spaces, a simple dimensionality reduction algorithm was implemented.

It begins by filtering all patents by a composed query, and placing those that pass the filter into random locations in a two dimensional space. The filter currently passes all documents that match every term in the user's query, but passing an N-best set of results from a weighted combination of any of the distance measures would also be appropriate. The randomly-placed documents are then iteratively 'pushed' in a way that locally minimizes the disparity between their high-dimensional and two dimensional distances, until they converge to a local minimum or a user-specified number of iterations runs.

In order to make the algorithm converge quickly while avoiding overshoot problems in which nodes are pushed to a higher distance disparity, an acceleration and backoff system was implemented. The rate at which nodes move toward and away from one another is controlled by a velocity parameter, the initial value of which was determined experimentally. For each iteration in which the distance disparity is lower than in the previous iteration, the velocity increases by a small multiplicative factor. In cases where the strain has increased – suggesting an overshoot condition–the velocity decreases exponentially more rapidly with each consecutive round of overshoot.

The dimensionality reduction algorithm uses the pairwise distance between each pair of points for hundreds of iterations, so these distances are cached. As a result, the first iteration is the only one that takes a substantial amount of time.

3.5 Clustering

Supposing the dimensionality reduction finds a local minimum that is reasonably close to the global minimum, patents that are similar in the original feature space are displayed as clusters in the two-dimensional representation. Patents are then clustered in an effort to make the visual presentation of patents more clear and to allow patents that are extremely similar to be retrieved together, such as those that are slightly different variations on the same template.

The clustering approach used is a simple bottom-up, pairwise algorithm, grouping patents or groups of patents by their centroids until certain criteria are met, such as the minimum distance crossing a user-determined threshold. Because a single title no longer serves as a short descriptor for nodes for visualization purposes, a short summary of multiple-patent clusters is computed by picking out the most significant words in all patents in a cluster by log-entropy as a replacement.

3.6 Explanation system

Across all of the distance measures, there is a shared framework for explaining the basis of the similarity between documents. Each explanation generated for a pair of documents comprises a similarity value and a listing of components contributing to the similarity, sorted by the extent to which they suggest the documents are similar.

3.6.1 Lexical Distance Explanation

In the lexical space, the components that contribute most to the similarity of two documents are found by determining which words would most quickly make the documents more similar given a proportional incremental increase in their word significances. For instance, if a word occurs on only one document, then making it more significant would render two documents less similar. These per-component contributions are found by taking the partial derivative of the cosine between two documents with respect to the weight of every word:

$$\frac{2W_k X_{AK} X_{BK}}{|X_A|_w |X_B|_w} - \left\langle X_A, X_B \right\rangle \frac{Wk(|X_B|_w^2 \; X_{AK} + |X_A|_w^2 \; X_{BK})}{(|X_A|_w |X_B|_w)^3}$$

Where $|X_A|_w$ is the norm of the weighted vector A, equivalent to $|X_A|$ given weights equal to one.

This derivative has the additional benefit of laying the groundwork for allowing users to tell the system to make two documents more similar in a way that may improve overall clustering and retrieval purpose. In traversing a weight gradient to make two documents more similar, the system could change the global significance of features and weights semantically useful terms more heavily.

3.6.2 LSA Distance Explanation

A distance explanation exactly parallel to the lexical explanation was tried for LSA, in which the weighted cosine derivative was calculated for components and the components were summarized along with the terms that made them fit the documents being compared. As mentioned before, these components are sometimes hard to make sense of, to the point where the explanations provided were useless. In response, a less direct explanation tool was built that finds synonym sets for the words appearing in the documents, and matches the documents by finding synset overlaps. The weight of a given component relating document A and document B in this scheme is the weight of a word in document A times the extent to which A is related to a synonym for a word in document B, times the extent that the synonym relates to the word in B, times the weight of a word in B. In general, the top components are words that are shared between the two documents, that is, the synonym is the word itself in both cases. These are followed by words in A that have synonyms related to words in B, and vice versa, followed by word pairs that are linked by non-identical synonyms.

As an example, the summary for patent numbers 6387375 and 6444872, which are titled “Methods and compositions for controlling coleopteran infestations” and “Large animal model of invasive pulmonary aspergillosis in an immunocompromised host”. The first patents the use of a fungus to kill insects, and the second patents a technique to suppress the immune system of animals in order to infect them with a fungal lung disease.

```
< 0.38975790093066887  PAR: fungus      C0: fusarium C1: aspergillus >
< 0.14165439708504668  PAR: species     C0: species  C1: species >
< 0.11737733991808748  PAR: host        C0: host     C1: immunocompromised >
< 0.06316592589532023  PAR: radiation   C0: spore    C1: irradiation >
< 0.05026810036688147  PAR: composition C0: composition      C1: comprising >
< 0.04564482390997831  PAR: number      C0: accession        C1: lobe >
< 0.03373284273941739  PAR: said        C0: claim    C1: said >
< 0.027545158303756343 PAR: fungus      C0: nrrl     C1: antifungal >
< 0.019844074797226184 PAR: group       C0: group    C1: consisting >
< 0.00899483513085717  PAR: polypeptide C0: comprising       C1: appetite >
< 0.005929008508094303 PAR: control     C0: control  C1: early >
< 0.005825546187228575 PAR: recombinant C0: purified C1: established >
< 0.004601196143206829 PAR: combination C0: nos      C1: combination >
< 0.002572620668014799 PAR: plant       C0: plants   C1: uninfected >
< 0.001385439166577634 PAR: group       C0: selected C1: claim >
```

The terms that capture the similarity of the documents most directlyare the species of fungus that the patents describe, and they are linked through the common synonym ‘fungus’. The next component is ‘species’, which receives a perfect relatedness score but a low word significance score. As the list continues the connections become more tenuous and the words less significant.

It is hoped that a larger basis matrix for the Singular Value Decomposition will allow the first approach to produce results that are intuitive to the user.

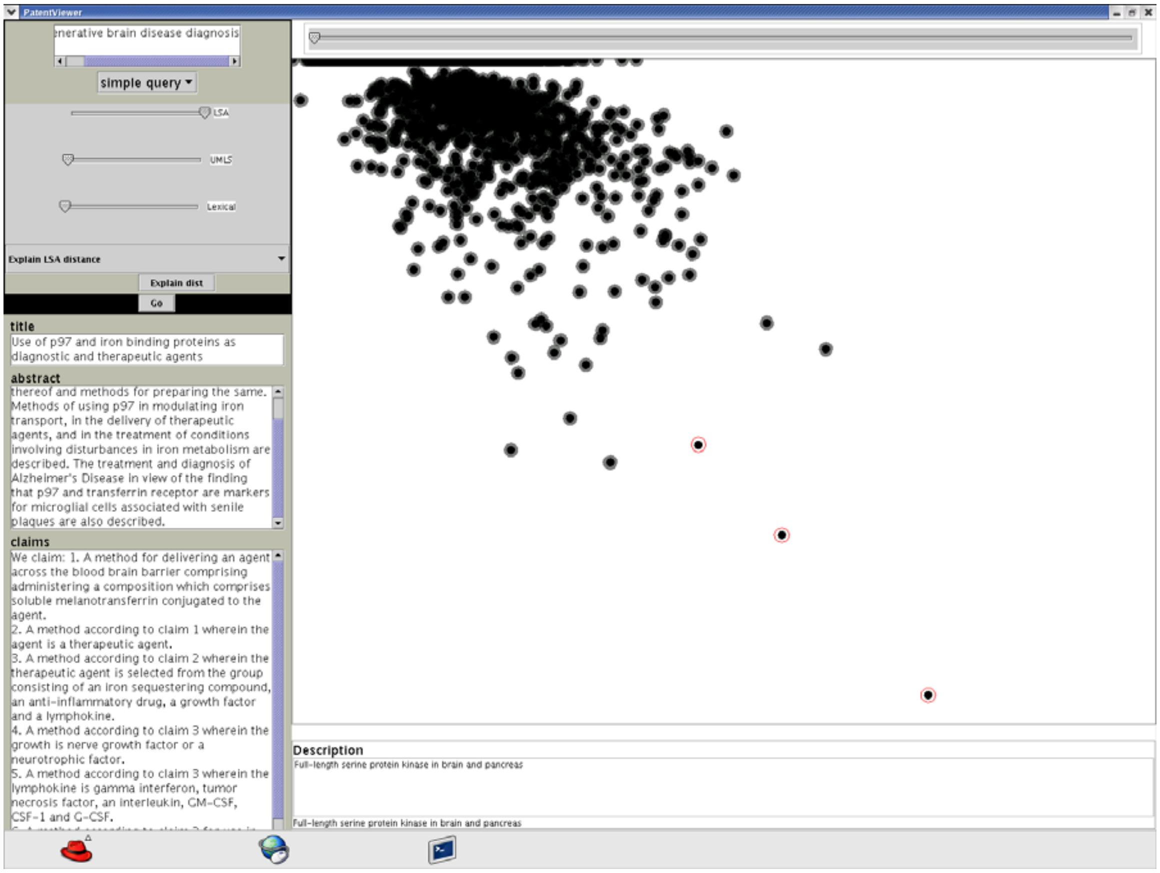
3.6.3 JC Distance Explanation

The Jiang-Conrath explanation system lists the best matches between the concepts in the documents or document and query being compared, starting with the exact matches.

4. USER INTERFACE

This section is a walkthrough of the user interface using a fictional example.

Suppose a patent attorney is working on behalf of an inventor who has recently developed a genetics-based method to diagnose a family of degenerative brain disorders. The first thing the attorney needs to do is form a query and retrieve documents with it. The first thing he does is compose a query in the top left text area. Directly below the text area is a drop-down menu that lets the user choose to perform either straight retrieval or dimensionality reduction and clustering. Below that are the sliders for the weightings of the different distance measures. The default behavior is to perform straight retrieval using LSA distance only, which is what the attorney will do first, with the query "degenerative brain disease diagnosis". In about two seconds, the system returns a scatter plot of the documents closest to the query from right to left. There are two dimensions available for visualization and there is only one dimension to a simple query, so the system fills in the additional dimension with lexical differences. It can alternately be configured to randomly assign y-values.



enerative brain disease diagnosis

title

Use of p97 and iron binding proteins as diagnostic and therapeutic agents

abstract

thereof and methods for preparing the same. Methods of using p97 in modulating iron transport, in the delivery of therapeutic agents, and in the treatment of conditions involving disturbances in iron metabolism are described. The treatment and diagnosis of Alzheimer's Disease in view of the finding that p97 and transferrin receptor are markers for microglial cells associated with senile plaques are also described.

claims

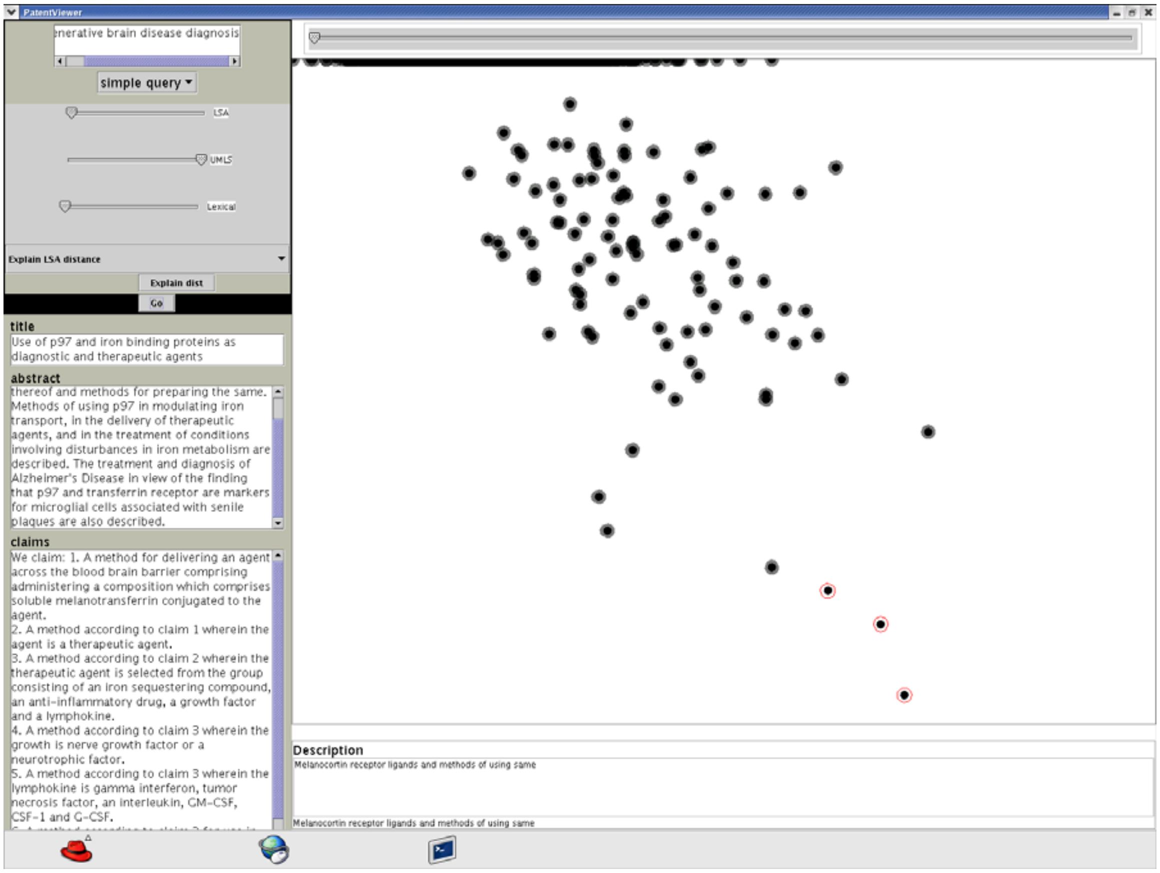
We claim: 1. A method for delivering an agent across the blood brain barrier comprising administering a composition which comprises soluble melanotransferrin conjugated to the agent.
2. A method according to claim 1 wherein the agent is a therapeutic agent.
3. A method according to claim 2 wherein the therapeutic agent is selected from the group consisting of an iron sequestering compound, an anti-inflammatory drug, a growth factor and a lymphokine.
4. A method according to claim 3 wherein the growth is nerve growth factor or a neurotrophic factor.
5. A method according to claim 3 wherein the lymphokine is gamma interferon, tumor necrosis factor, an interleukin, GM-CSF, CSF-1 and G-CSF.

Description

Full-length serine protein kinase in brain and pancreas

Full-length serine protein kinase in brain and pancreas

There are several appropriate hits to be found using a quick scan of the summary panel at the bottom of the screen while placing the cursor near the best-fitting results, including “Methods for the prevention or treatment of alzheimer’s disease”, “Diagnostic methods for alzheimer’s disease by detection of multiple mRNAs”, and “In vitro assay method for the study of brain aging”. The attorney can highlight these patents, giving them a red outline, and compare the LSA results to Jiang-Conrath results by running a JC query, show below. Note the relatively sparse coverage.



nerative brain disease diagnosis

title

Use of p97 and iron binding proteins as diagnostic and therapeutic agents

abstract

thereof and methods for preparing the same. Methods of using p97 in modulating iron transport, in the delivery of therapeutic agents, and in the treatment of conditions involving disturbances in iron metabolism are described. The treatment and diagnosis of Alzheimer's Disease in view of the finding that p97 and transferrin receptor are markers for microglial cells associated with senile plaques are also described.

claims

We claim: 1. A method for delivering an agent across the blood brain barrier comprising administering a composition which comprises soluble melanotransferrin conjugated to the agent.

2. A method according to claim 1 wherein the agent is a therapeutic agent.

3. A method according to claim 2 wherein the therapeutic agent is selected from the group consisting of an iron sequestering compound, an anti-inflammatory drug, a growth factor and a lymphokine.

4. A method according to claim 3 wherein the growth is nerve growth factor or a neurotrophic factor.

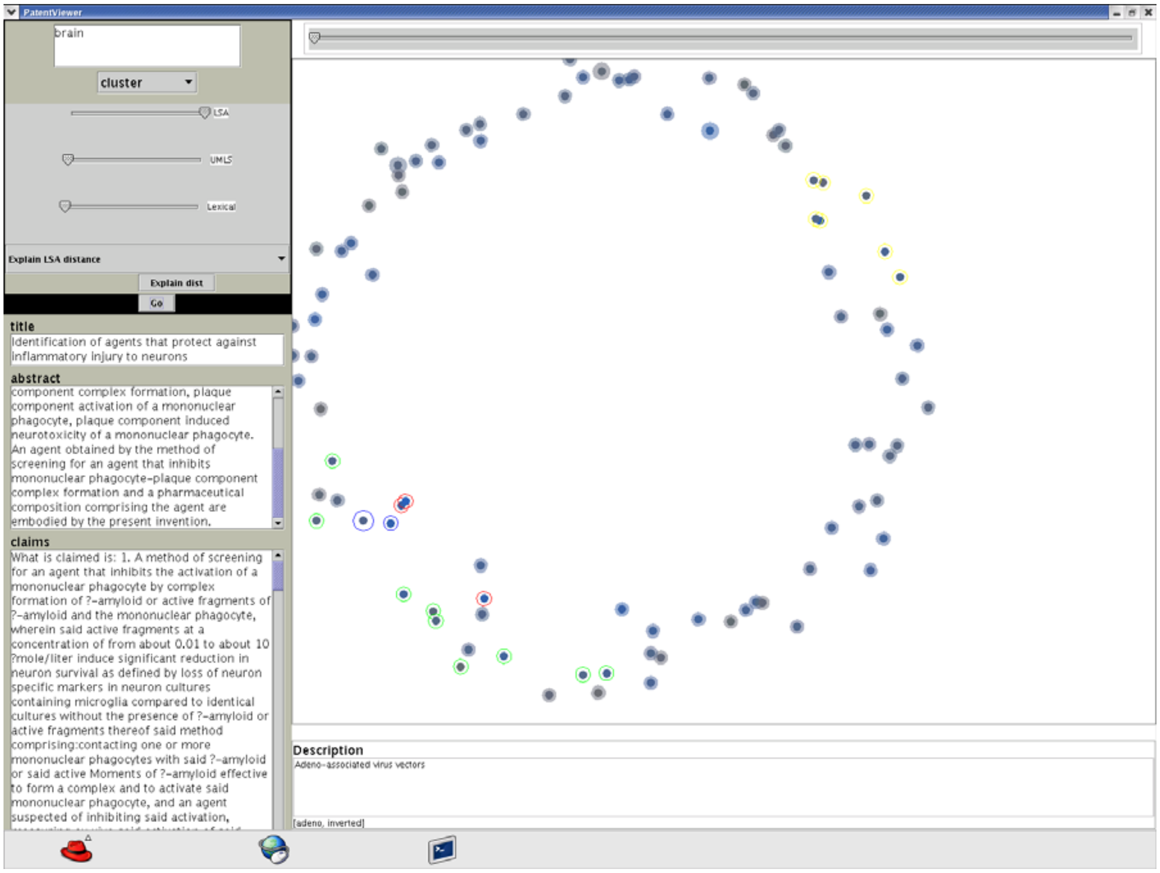
5. A method according to claim 3 wherein the lymphokine is gamma interferon, tumor necrosis factor, an interleukin, GM-CSF, CSF-1 and G-CSF.

Description

Melanocortin receptor ligands and methods of using same

Melanocortin receptor ligands and methods of using same

If the attorney wants to see what other kinds of patents might be related to these three and possibly discover some other directly relevant patents, he can use clustering. Because clustering filters by the query, he trims it to simply “brain”. In the LSA clustering screen, additional patents that have to do with diagnosis of degenerative brain disease including “method for the diagnosis of brain/neurological disease using monoclonal antibodies” have been labeled in green. Patents having to do with the study and treatment of the neurodegenerative diseases have been labeled in blue, and include “rodent models of human amyloidoses”, the large node which denotes a cluster of three patents which have the same title, “Identification of agents that protect against inflammatory injury to neurons” but somewhat different claims. The attorney can click on this node or any other and bring up a list of member patents to read in detail. The full claims of patents thus selected appear in the lower right set of text boxes. To provide random example of the coherence of clusters, a group of patents all having to do with gene delivery have been highlighted in yellow. The circular arrangement of the nodes is due to the normalization of the LSA distance giving unrelated patents a distance of 0.5.



brain

title

Identification of agents that protect against inflammatory injury to neurons

abstract

component complex formation, plaque component activation of a mononuclear phagocyte, plaque component induced neurotoxicity of a mononuclear phagocyte. An agent obtained by the method of screening for an agent that inhibits mononuclear phagocyte-plaque component complex formation and a pharmaceutical composition comprising the agent are embodied by the present invention.

claims

What is claimed is: 1. A method of screening for an agent that inhibits the activation of a mononuclear phagocyte by complex formation of ?-amyloid or active fragments of ?-amyloid and the mononuclear phagocyte, wherein said active fragments at a concentration of from about 0.01 to about 10 ?mole/liter induce significant reduction in neuron survival as defined by loss of neuron specific markers in neuron cultures containing microglia compared to identical cultures without the presence of ?-amyloid or active fragments thereof said method comprising:contacting one or more mononuclear phagocytes with said ?-amyloid or said active Moments of ?-amyloid effective to form a complex and to activate said mononuclear phagocyte, and an agent suspected of inhibiting said activation,

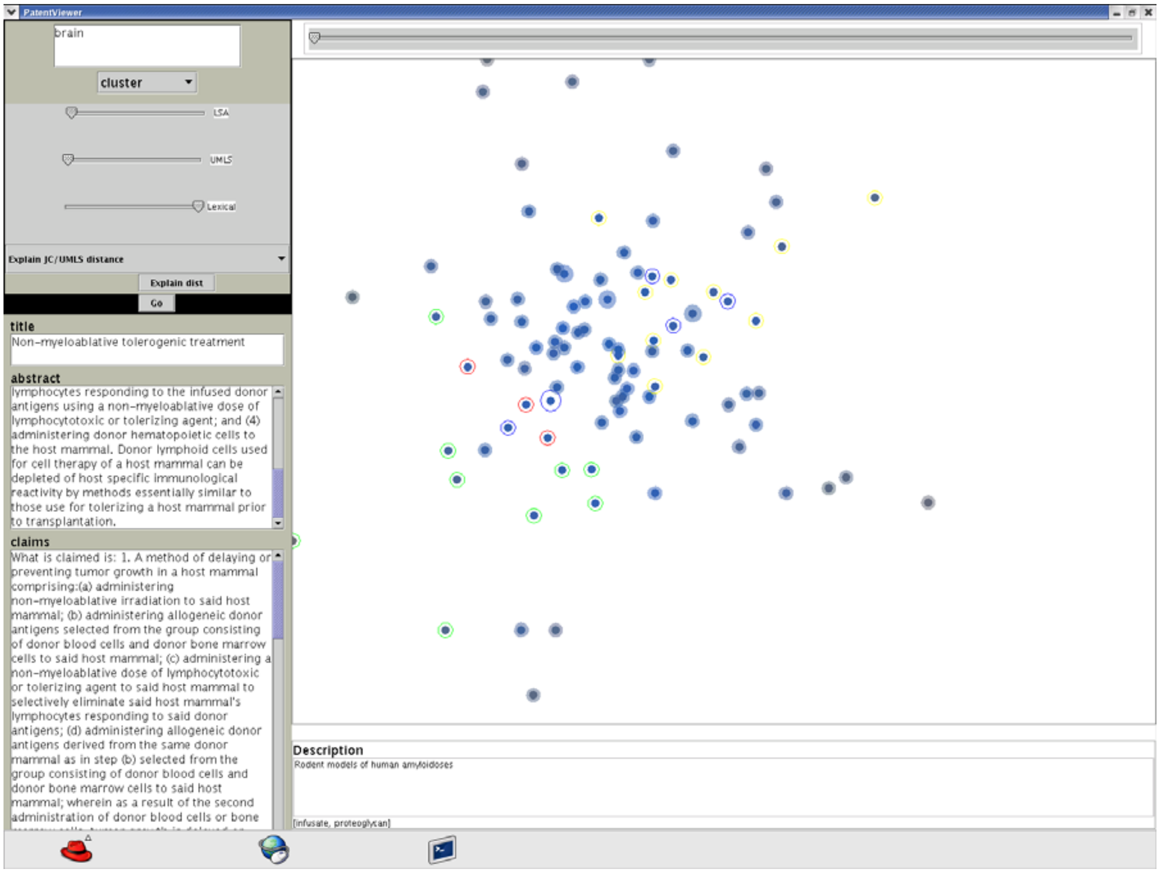
Description

Adeno-associated virus vectors

[adeno, inverted]

The next screen shot shows lexically-driven clustering with the same labeled points, an additional set of three gene-delivery related patents that has been highlighted blue amidst those that were labeled yellow.

In the neurodegenerative disease diagnosis group there is an unselected patent, which has the title “Assay and reagent kit for evaluation of multi-drug resistance in cells” which might give the attorney some pause.



brain

cluster

Explain JC/UMLS distance

title

Non–myeloablative tolerogenic treatment

abstract

lymphocytes responding to the infused donor antigens using a non–myeloablative dose of lymphocytotoxic or tolerizing agent; and (4) administering donor hematopoietic cells to the host mammal. Donor lymphoid cells used for cell therapy of a host mammal can be depleted of host specific immunological reactivity by methods essentially similar to those use for tolerizing a host mammal prior to transplantation.

claims

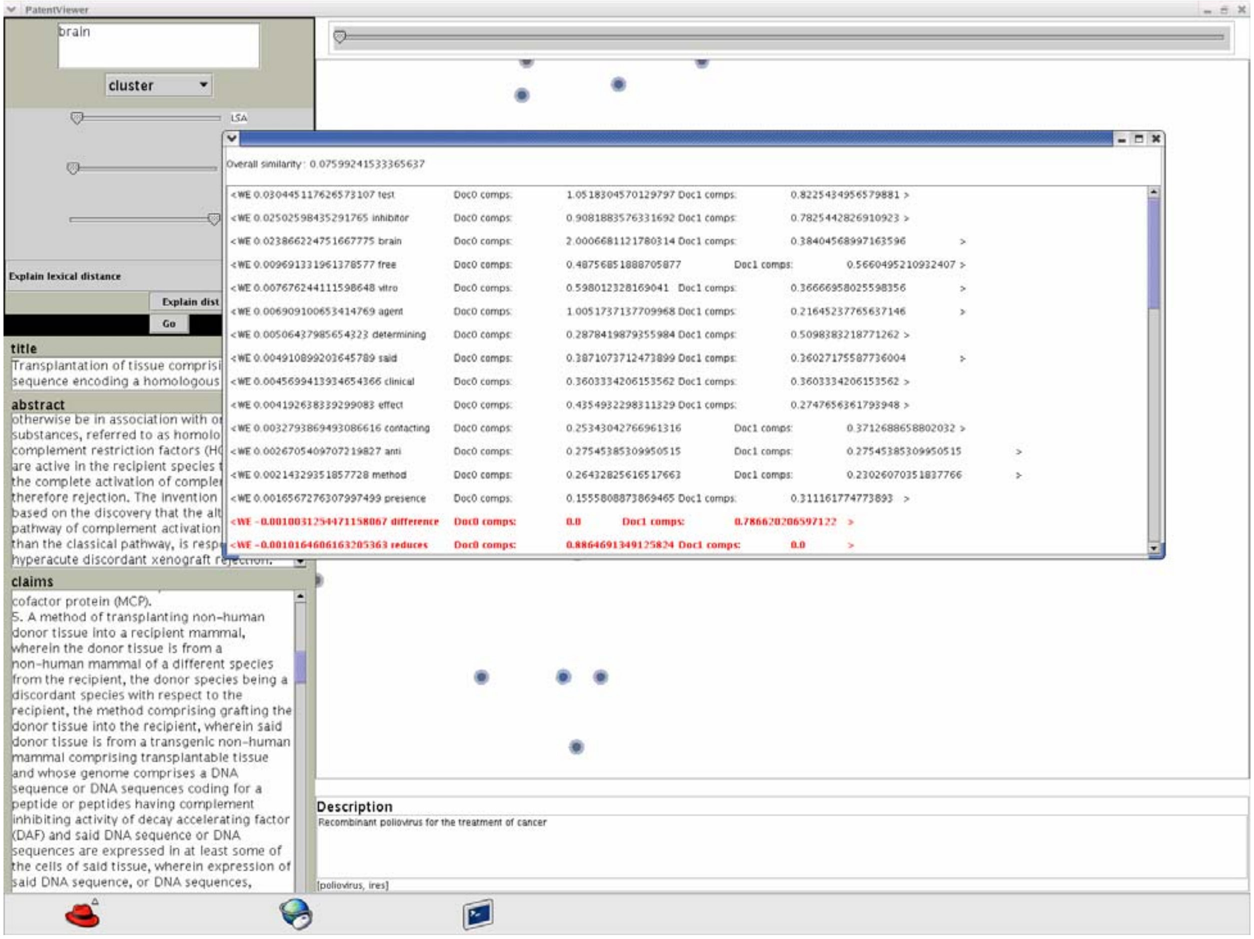
What is claimed is: 1. A method of delaying or preventing tumor growth in a host mammal comprising:(a) administering non–myeloablative irradiation to said host mammal; (b) administering allogeneic donor antigens selected from the group consisting of donor blood cells and donor bone marrow cells to said host mammal; (c) administering a non–myeloablative dose of lymphocytotoxic or tolerizing agent to said host mammal to selectively eliminate said host mammal's lymphocytes responding to said donor antigens; (d) administering allogeneic donor antigens derived from the same donor mammal as in step (b) selected from the group consisting of donor blood cells and donor bone marrow cells to said host mammal; wherein as a result of the second administration of donor blood cells or bone

Description

Rodent models of human amyloidoses

[infusate, proteoglycan]

In order to make sense of its placement, the attorney can use the explanation feature. Because the lexical basis was used for this clustering, it makes sense to use the lexical explanation basis. Comparing this to one of the original best matches reveals that it is considered similar because the documents have similar proportions of the words test, inhibitor, brain, free, vitro, agent, and determining, and their difference is largely due to the bad example lacking the words aging, symptom, slice, cortex, and neocortex. Some of these latter words might be used to improve retrieval.



title
Transplantation of tissue comprisi
sequence encoding a homologous

abstract
otherwise be in association with o
substances, referred to as homolo
complement restriction factors (HC
are active in the recipient species t
the complete activation of comple
therefore rejection. The invention
based on the discovery that the alt
pathway of complement activation
than the classical pathway, is resp
hyperacute discordant xenograft rejection.

claims
cofactor protein (MCP).
5. A method of transplanting non-human donor tissue into a recipient mammal, wherein the donor tissue is from a non-human mammal of a different species from the recipient, the donor species being a discordant species with respect to the recipient, the method comprising grafting the donor tissue into the recipient, wherein said donor tissue is from a transgenic non-human mammal comprising transplantable tissue and whose genome comprises a DNA sequence or DNA sequences coding for a peptide or peptides having complement inhibiting activity of decay accelerating factor (DAF) and said DNA sequence or DNA sequences are expressed in at least some of the cells of said tissue, wherein expression of said DNA sequence, or DNA sequences,

```
Overall similarity : 0.07599241533365637

<WE 0.030445117626573107 test           Doc0 comps:   1.0518304570129797 Doc1 comps:   0.8225434956579881 >
<WE 0.02502598435291765 inhibitor       Doc0 comps:   0.9081883576331692 Doc1 comps:   0.7825442826910923 >
<WE 0.02386622475166775 brain           Doc0 comps:   2.0006681121780314 Doc1 comps:   0.38404568997163596         >
<WE 0.009691331961378577 free           Doc0 comps:   0.48756851888705877          Doc1 comps:   0.5660495210932407 >
<WE 0.007676244111598648 vitro          Doc0 comps:   0.598012328169041  Doc1 comps:   0.36666958025598356         >
<WE 0.006909100653414769 agent          Doc0 comps:   1.0051737137709968 Doc1 comps:   0.21645237765637146         >
<WE 0.00506437985654323 determining    Doc0 comps:   0.2878419879355984 Doc1 comps:   0.5098383218771262 >
<WE 0.004910899203645789 said           Doc0 comps:   0.3871073712473899 Doc1 comps:   0.36027175587736004         >
<WE 0.0045699413934654366 clinical     Doc0 comps:   0.3603334206153562 Doc1 comps:   0.3603334206153562 >
<WE 0.004192638339299083 effect         Doc0 comps:   0.4354932298311329 Doc1 comps:   0.2747656361793948 >
<WE 0.0032793869493086616 contacting   Doc0 comps:   0.25343042766961316          Doc1 comps:   0.3712688658802032 >
<WE 0.0026705409707219827 anti         Doc0 comps:   0.27545385309950515          Doc1 comps:   0.27545385309950515         >
<WE 0.00214329351857728 method         Doc0 comps:   0.26432825616517663          Doc1 comps:   0.23026070351837766         >
<WE 0.0016567276307997499 presence     Doc0 comps:   0.1555808873869465 Doc1 comps:   0.311161774773893  >
<WE -0.0010031254471158067 difference   Doc0 comps:   0.0          Doc1 comps:   0.786620206597122  >
<WE -0.0010164606163205363 reduces      Doc0 comps:   0.8864691349125824 Doc1 comps:   0.0          >
```

Description
Recombinant poliovirus for the treatment of cancer
[poliovirus, ires]

Finally, the attorney might want to get a sense of the general terms capturing the commonalities between two patents. This can be done most directly by using the JC distance explanation. Looking at its output and excluding uninformative commonalities like the verb ‘is’, the attorney sees that the documents have concepts with common parents including “brain”, “nervous system diseases”, “cellular structures”, and “general concepts of health and disease”.

5. CONCLUSION

This thesis has described a system that can retrieve documents in a way that, unlike currently prevailing search technologies, allows users to see results in a richer context than merely goodness of fit, allows users to understand why the system organizes its results as it does, and gives users some control over their queries as a deeper level than stringing words together. It integrates multiple distance measures and features that give users different ways of looking at documents, and two-dimensional visualization, clustering and dimensionality reduction that allow them to seen relationships between documents and groups of documents. It approach has several limitations in its current form, and there as many unexplored directions in which it might be taken.

One aspect of the project that may be easily improved is the resolution at which users have control over searches. The current distance measures and their explanations allow users to manually modify their queries by content and by weight, but do not allow users to alter the weights underlying the distance measures themselves. The implementation of manual and automatic weight adjustments through the explanation system would improve system performance and is consistent with the idea that a system to find and organize information should be able to modify its operation to suit the exact goals of the user, who, for example, may in one case want to draw a distinction between corn and soybeans, but in another case may care more about transgenic versus hybrid crops. All of the distance measures lend themselves to gradient descent of parameters to automatically maximize or minimize document similarities as the user sees fit, as well as user-directed changes in the weight given to specific components.

Another place where immediate improvement might be made in the project is in its dimensionality reduction algorithms. The approach currently in use is neither especially fast nor does it claim to escape local minima very well. In addition to exploring approaches that do a good job of globally aligning low and high-dimensional distances, it may be worthwhile to consider methods that pay special attention to locally correct distances, such as that described in (de Silva and Tenenbaum, 2002).

The coverage and appropriateness of the ontology are also significant issues, and addressing them may not be easy. Personal communication with Michael Collins suggested that an extension of some of his and Yoram Singer's work in named entity labeling (Collins and Singer, 1999) might permit relatively efficient extension of ontologies for specific problem domains.

With respect to the problems with mapping the LSA-based representations to human intuitions, the only immediate possibility is to try to wash out problems caused by noise by using much larger basis matrix.

Finally, the clustering method's time complexity, which is $O(kN^2)$ where N is the number of nodes and k is the number of iterations, is a bottleneck in increasing data set size used in clustering. Using commonly known algorithms for computing pairwise distance and caching, it may be improved to $O(N \log N)$ time.

The amount of potentially useful text stored in databases and collections will continue to increase. One consequence of this is that search and pattern discovery systems will be applied by an ever-growing number of organizations with diverse goals. Systems that are specifically satisfy some goals – like web search - will be poorly suited to other goals. In light of this, there appear to be two possible directions for the technology behind search and visualization of data. The first involves an endless parade of purpose-specific tools, developed as new needs arise. The Patent Semantics system is a step in a second direction, toward systems that provide organizations with such diverse and intuitive ways of organizing data that new needs can be met as they arise.

Appendix A – Database technical details

Citations table (patents.discits)

- Maps citing patent numbers to cited patent numbers.

```
+----------+---------+------+-----+---------+-------+
| Field    | Type    | Null | Key | Default | Extra |
+----------+---------+------+-----+---------+-------+
| pnum     | int(11) | YES  |     | NULL    |       |
| citation | int(11) | YES  |     | NULL    |       |
+----------+---------+------+-----+---------+-------+
```

Patents table (patents.dispats)

- One entry for each patent.
- Contains patent number, title, abstract, claims, issue date, and assignee.

```
+----------+---------+------+-----+---------+-------+
| Field    | Type    | Null | Key | Default | Extra |
+----------+---------+------+-----+---------+-------+
| pnum     | int(11) | YES  |     | NULL    |       |
| title    | text    | YES  |     | NULL    |       |
| abstract | text    | YES  |     | NULL    |       |
| claims   | text    | YES  |     | NULL    |       |
| date     | text    | YES  |     | NULL    |       |
| assignee | text    | YES  |     | NULL    |       |
+----------+---------+------+-----+---------+-------+
```

Phrases table (patents.disphr)

- Maps patent numbers to phrases.
- Contains patent numbers, phrase locations (ABS/CLM), phrases' CUIs in UMLS or '??', and phrase text.

```
+----------+-------------+------+-----+---------+-------+
| Field    | Type        | Null | Key | Default | Extra |
+----------+-------------+------+-----+---------+-------+
| pnum     | int(11)     | YES  |     | NULL    |       |
| location | varchar(16) | YES  |     | NULL    |       |
| type     | varchar(16) | YES  |     | NULL    |       |
| cui      | varchar(16) | YES  |     | NULL    |       |
| contents | text        | YES  |     | NULL    |       |
+----------+-------------+------+-----+---------+-------+
```

Phrases table (patents.disphr)

- Maps patent numbers to classifications.
- Contains patent number and US classification (423_###)

```
+----------------+-------------+------+-----+---------+-------+
| Field          | Type        | Null | Key | Default | Extra |
+----------------+-------------+------+-----+---------+-------+
| pnum           | int(11)     | YES  |     | NULL    |       |
| classification | varchar(16) | YES  |     | NULL    |       |
+----------------+-------------+------+-----+---------+-------+
```

Appendix B – Natural Language Processing Details

The Brill tagger and Collins parser were trained using the Penn TreeBank corpus of manually tagged Wall Street Journal text. Unfortunately, the vocabulary and syntax of patent abstracts and claims are idiosyncratic and quite unlike those found in newspaper text. An effort was made to annotate a small corpus of patent text to improve performance, but it was halted due to personnel limitations.

Tagging and Parsing problems and workarounds:

First, there were non-ASCII characters that caused the parser to crash. These characters were thrown out wholesale, such that ‘tΘast’ would become ‘tast’.

Problems at the word level were dominated by mistagging. Out of vocabulary words were especially likely to be mistagged. For example, the nouns mistagged as verbs included ‘triphosphates’, ‘oligonucleotide’, and ‘streptavidin’ – words that appear rarely, if at all, in newspaper text.

Another problem with individual words is that some chemical names in patents are very long, sometimes over a hundred characters. Very long words caused buffer overflow problems in the tagger and were simply thrown out. Given more time, a more principled option would have been to replace these words with unique, short, non-word identifiers and maintain a mapping such that they could be mapped back to their original strings in a post processing step.

There were also problems with sentence length – the parser failed on extremely long sentences of over four hundred words which occurred in the patent data.

A tiered approach was taken to keeping sentence length manageable. All sentences were first divided into smaller sentences based on semicolon and colon boundaries. All sufficiently short resultant sentences were parsed, and those that were too still too long were segmented based on commas. This did not result in grammatical sentences, but given the goal of mining noun and verb phrases, the resultant fragments were still useful. Any fragments that were still too long to parse were then thrown out.

There is no doubt that the amount of potentially useful text stored in databases and collections will continue to increase. One consequence of this is that search and pattern discovery systems will be applied by an ever-growing number of organizations with diverse goals. Systems that are specifically designed to satisfy some goals – like web search - will be poorly suited to other goals. In light of this, there appear to be two possible directions for the technology behind search and visualization of data. The first involves an endless parade of purpose-specific tools, developed as new needs arise, with the consequences of poor economies of scale for software development and support, and high training costs. The Patent Semantics system is a step in a second direction, toward systems that provide organizations with such diverse and intuitive ways of organizing data that new needs can be met as they arise.

Appendix C – UMLS Data Structures

In order to extract ancestry relations between concepts that are necessary to compute Jiang-Conrath distance, the UMLS Concept contexts (umlsMT.MRCXT) table was used. Specifically, one concept was judged to have another as an ancestor if and only if there existed an ancestor (ANC) relation between the two in the context column (CXT) of the table. There are other ways to obtain ancestry relations, and this was used for two reasons:

- The 'isa' and other more precise semantic relations have poorer coverage than ANC relations. The cost of imprecision was judged to be lower than the cost of very sparse coverage.
- The organization of the ancestor data made it possible to compute Jiang-Conrath distances relatively quickly without building massive cached data structures.

Given that the coverage of the UMLS is increasing yearly, in the future it may be wise to use the more specific semantic relations in the UMLS Semantic Net.

The current approach leads to certain problems due to the structure of the ancestry relations. In some cases, concepts are their own ancestors. This problem is addressed by a simple check during lookup. In other cases, the counts of some concepts are greater than those of their ancestors, leading occasionally to distances between concepts that are less than zero. This issue is dealt with by setting illegal negative distances to fixed positive values.

Appendix D – Source code organization

This appendix lists the java classes that are using in the Patent Semantics System, and their function and initialization where nontrivial.

Parsing patent from USPTO into database

clucas.readPatents.PatentXMLHandler

- Parses XML files provided by the US Patent office, invokes natural language processing tools, attaches UMLS concepts to phrases, and populates patents database. Takes paths of XML files as arguments.

Retrieving data from database

clucas.tools.dbtools.DBManager

- Mediates all interaction with SQL database.
- Initialized with IP address of database server and database name. Username and password are hard-coded.

clucas.tools.dbtools.NewPatDocs

- Provides an abstraction layer on top on patent database.

- Initialized with a Map containing query parameters and a character mapping file which is null by default, or using static factory method NewPatDocs.limitNPatDocs(int n).

Preprocessing:

clucas.lsa.Preprocessor

- Computes and stores all (1-entropy) weights for words, filtering by a stop word list.
- Initialized with an instance of NewPatDocs, a minimal document frequency for a word to be considered, and the path to a stop word file.

clucas.lsa.StopWords

- Reads a stop word file and returns boolean function determining stop word status of words.

Distance measures and explanation

General explanation interfaces:

clucas.ui.Explanation

clucas.ui.ExplanationComponent

- Interfaces providing common methods for Explanation tools to implement.

Lexical distance and explanation:

clucas.ui.DPDistanceGrabber

- Handles all lexical distance and explanation.
- Initialized with weighted vocabulary from clucas.lsa.Preprocessor.

Jiang-Conrath distance and explanation:

clucas.umls.SimpleJCDocDist

- Handles all document-level Jiang-Conrath distance and explanation.
- Initialized with an instance of clucas.umls.SimpleJCDist

clucas.umls.SimpleJCDist

- Implements the Jiang-Conrath algorithm using the umlsMT database.
- Initialized with an instance of clucas.tools.dbtools.NewPatDocs, a path to serialized count files, and an instance of clucas.tools.texttools.Morph.

clucas.umls.UmlsManager

- Provides an abstraction layer on top of the umlsMT database.
- Initialized with an optionally null String restricting the portion of the UMLS Metathesaurus considered.

clucas.umls.SimpleMatch

- Data structure storing explanation data for Jiang-Conrath distances.

LSA distance and explanation:

clucas.lsa.LsaDocDist

- Handles all LSA distance and explanation.
- Initialized with an instance of clucas.lsa.LsaCore

clucas.lsa.LsaCore

- Implements LSA algorithms and stores data structures.
- Initialized with path to serialized version, or paths to tab-delimited matrices. Example initialization code is in clucas.Lsa.LsaRunner.

clucas.lsa.LsaRunner

- Creates, reads, and tests LsaCore.
- Initialized with path to serialized LsaCore or target LsaCore location and target number of dimensions.

clucas.lsa.ComponentExplanation

clucas.lsa.LsaExplanation

- Data structures storing explanation data for LSA distances.

User interface and dispatching code:

clucas.ui.PatentViewer

- Starting execution point for Patent Semantics System.
- Takes configuration file as argument.

clucas.ui.SpaceManager

- Handles sizing for interface components.

clucas.ui.SpaceViewer

- Handles zooming and panning of MainVisPanel

clucas.ui.QueryPanel

- Contains components for writing and executing queries.

clucas.ui.GadgetPanel

- Contains components for changing query weights and selecting explanation types.

clucas.ui.ProgBar

- Reports dimensionality reduction and clustering performance.

clucas.ui.MiniView

- Displays patent titles and summaries.

clucas.ui.ConfigReader

- Reads configuation file to populate clucas.ui.Constants.

clucas.ui.Constants

- Clearinghouse for global configuration variables.

clucas.ui.ExplainPane

- Displays distance explanations.

clucas.ui.SearchManager

- Dispatches queries to backend tools.

clucas.ui.ClusterQueryResults

- Data structure for shuttling results to viewer.

clucas.ui.PostIt

- Displays patents contained in nodes for fetching full text fields.

clucas.ui.MainVisPanel

- Draws visualization of results.

clucas.ui.PatentDisplayPanel

- Displays title, abstract, and claims of selected patents.

Clustering and dimensionality reduction:

clucas.cluster.Anneal

- Implements dimensionality reduction and clustering algorithms. Initialized by clucas.ui.SearchManager.

clucas.cluster.AnnealPoint

- Data structure for use in dimensionality reduction and clustering.

Miscellaneous math functions, data structures, and IO

clucas.tools.VectorTools

clucas.tools.CountedCollection

clucas.tools.WeightedCollection

clucas.tools.Multimap

clucas.tools.MatrixWriter

clucas.tools.Log

BIBLIOGRAPHY

Baker, C., Fillmore, C., and Lowe, J. (1998). The Berkeley FrameNet Project. In *Proceedings of COLING-ACL '98*. (pp. 86-90). Association for Computational Linguistics.

Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92*, (pp. 152–155).

Brin, S. and Page, L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems,* 30(1-7):107-117, April 1998.

Budanitsky, A. and Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Collins, M. and Singer, Y. (1999). Unsupervised Models for Named Entity Classification. EMNLP/VLC-99.

Dumais, S. T. (1991), Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers, 23(2)*, 229-236.

Fellbaum, C. editor. 1998. ***WordNet:*** *An Electronic Lexical Database*. The MIT Press.

Gildea, D. and Jurafsky, D. 2002. *Automatic labeling of semantic roles*. Computational Linguistics, 28(3):245—288

IPVision, Inc. Internet: http://www.tecpatents.com/ipvision/ipportal.htm. August 20, 2004.

Larkey, L. (1999) A Patent Search and Classification System. *Digital Libraries 99 -The Fourth ACM Conference on Digital Libraries* (Berkeley, CA, Aug. 11-14 1999) ACM Press, pp. 79-87.

Nelson, S., Powell, T., Humphreys, B. (2002). The Unified Medical Language System (UMLS) Project. In: Kent, Allen; Hall, Carolyn M., editors. *Encyclopedia of Library and Information Science.* New York: Marcel Dekker, Inc. p.369-378.

OmniViz – Applications. http://www.omniviz.com/applications/omni_viz.htm

August 20, 2004.

de Silva, V. and Tenenbaum, J. B. (2002). Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems 15.* M.SBecker, S., Thrun, S., and Obermayer, K. (eds). Cambridge, MIT Press, 2002, 705-712.