Distributed Object Search System and Method

Brief Summary

The present invention is a high-performance ontology-based search engine that supports the indexing of arbitrary objects, including images, sound and video streams, as well as traditional data objects such as text files and structured documents. Moreover, this indexing is by the actual content of the objects, not just the indexing of annotations or meta-data associated with the objects.

Abstract

A distributed computer system including one or more front end computers and one or more computer nodes interconnected by a network into a search engine for retrieval of objects including images, sound and video streams, as well as plain and structured text. A query is an object in the same format as the objects to be retrieved. A query from a user is transmitted to one of the front end computers which forwards the query to one of the computer nodes, termed the home node, of the search engine. The home node extracts features from the query and hashes these features. Each hashed feature is transmitted to one node on the network. Each node on the network which receives a hashed feature uses the hashed feature of the query to perform a search on its respective partition of the database. The results of the searches of the local databases are gathered by the home node. If requested, this process is repeated a second time by the home node to refine the results of the query.

Informal Discussion

Feature-vector and wavelet information retrieval. Algorithms are now commonly available for extracting feature vectors from images, audio, video and multimedia data. This patent shows how one can incorporate feature vectors into a traditional, text-based, natural language ontology, and how such an ontology can be used for ontology-based information retrieval using the jarg search engine.

Wavelets are a technique for efficient encoding of images. Other kinds of continuous data, such as audio, video and multimedia data can also be analyzed using wavelets. Many features of images are a function of the wavelet transformation, and the result of a discrete wavelet transformation is a sparse feature vector, so that wavelet analysis can be regarded as a special case of general feature extraction. The only distinction is that the sparse vector obtained by a generic transformation, such as a wavelet or Fourier transformation, does not have coefficients that have meaning in themselves. However, the sparse vector obtained by such a transformation can be used for image matching.

Although the obvious choice for the similarity function is the Euclidean distance, such a function does correspond to how people compute similarity. A more accurate similarity function has positive contributions for features that are in common, and negative contributions from features that are not in common.

The basic level of service of jarg will take care of features that are in common, as well as features of the query that are not in the indexed object, but it cannot handle features of the indexed object that are not in the query. For this one must use a higher level of service. This higher level of service can provide for

other attributes of the object, such as its URL and auxiliary data such as an expiration date.

Business Strategy

The main purpose of this patent is to clarify some ambiguities in the original jarg patent. That patent emphasized text and attribute based indexing. To index other kinds of objects, it proposed using content labels. Since the time of the original jarg patent, it has become possible to index objects directly using a variety of feature extraction techniques. Although one could argue that the feature extraction techniques are simply automating the construction of the content label, the patent might be construed as not being applicable for content-based indexing.

A secondary purpose of the patent is to clarify the higher levels of service introduced somewhat vaguely in the original patent. Because of this, it is possible for someone to file a patent for a search engine based on jarg but adding their own higher level of service. To prevent this, two specific higher levels of service were added to the claims.

Another secondary purpose of the patent is to introduce more kinds of similarity function than just the cosine. This is important because the cosine measure is primarily used for text retrieval and does not serve other kinds of retrieval as well. Although the original patent is not limited to a specific form of measure, it also does not mention how other the other similarity measures are related to the level of service. This means that the similarity measure is not just some function or algorithm applied at one point in the retrieval process, it is actually related to the architecture of the search engine. Since it is the architecture (or "manufacturing process") which is being patented, it is important to make any claims that are architecturally related to prevent someone else from filing a patent on it.

Metaphors

The original jarg patent was inspired by a technique in biochemistry for studying the genome in which small DNA sequences, called *probes*, are used for finding the locations of genes on a very long DNA sequence, such as a chromosome. The hashed fragments of jarg are analogous to biochemical probes, and the corpus of jarg is analogous to the genome. The present invention is even more strongly analogous to the biochemical technique, since the present invention indexes objects in general rather than flat text documents in particular.

Another inspiration for the original jarg patent was human memory and cognition. The precise mechanism of human memory and cognition is not yet known, but some aspects of it are well known. For example, our retrieval process is based on small aspects (or fragments) of larger memory units. As another example, damage to a person’s brain can make it more difficult to retrieve certain memories, but they are not completely lost.