A Knowledge Correlation Search Engine

Technical White Paper

Mark Bobick and W.Reid Cornwell Ph.D.

Search engines are widely acknowledged to be part of the Information Retrieval (IR) domain of knowledge. IR methods are directed to locating resources (typically documents) that are relevant to a question called a query. That query can take forms ranging from a single search term to a complex sentence composed in a natural language such as English. The collection of potential resources that are searched is called a corpus (body), and different techniques have been developed to search each type of corpus. For example, techniques used to search the set of articles contained in a digitized encyclopedia differ from the techniques used by a web search engine. Regardless of the techniques utilized, the core issue in IR is relevance – that is, the relevance of the documents retrieved to the original query. Formal metrics are applied to compare the effectiveness of the various IR methods. Common IR effectiveness metrics include precision, which is the proportion of relevant documents retrieved to all retrieved documents; recall, which is the proportion of relevant documents retrieved to all relevant documents in the corpus; and fall-out, which is the proportion of irrelevant documents retrieved to all irrelevant documents in the corpus. Post retrieval, documents deemed relevant are (in most IR systems) assigned a relevance rank, again using a variety of techniques, and results are returned. Although most commonly the query is submitted by – and the results returned to - a human being called a user, the user can be another software process.

Text retrieval is a type of IR that is typically concerned with locating relevant documents which are composed of text, and document retrieval is concerned with locating specific fragments of text documents, particularly those documents composed of unstructured (or "free") text.

The related knowledge domain of data retrieval differs from IR in that data retrieval is concerned with rapid, accurate retrieval of specific data items, such as records from a SQL database.

Information extraction (IE) is another type of IR which is has the purpose of automatic extraction of information from unstructured (usually text) documents into data structures such as a template of name/value pairs. From such templates, the information can subsequently correctly update or be inserted into a relational database.

The Knowledge Correlation Search Engine differs from existing search engines because the Knowledge Correlation process attempts to construct an exhaustive collection of paths describing all connections - called correlations - between one term, phrase, or concept referred to as X (or "origin") and a minimum of a second term, phrase or concept referred to as Y (or "destination"). If one or more such correlations can in fact be constructed, the Knowledge Correlation Search Engine identifies as relevant all resources which contributed to constructing the correlation(s). Unlike existing search engines, relevancy in the Knowledge Correlation Search Engine applies not to individual terms,

phrases or concepts in isolation but instead to the answer space of correlations that includes not only the X and the Y, but to all the terms, phrases and concepts encountered in constructing the correlations. Because of these novel characteristics, the Knowledge Correlation Search Engine is uniquely capable of satisfying user queries for which can not be answered using the content of a single web page or document.

Search engines that have been described in the literature or released as software products use a number of forms of input, ranging from individual keywords, to phrases, sentences, paragraphs, concepts and data objects. Although the meanings of *keyword*, *sentence*, and *paragraph* conform to the common understanding of the terms, the meanings of *phrase*, *concept*, and *data object* varies by implementation. Sometimes, the word *phrase* is defined using its traditional meaning in grammar. In this use, types of phrases include Prepositional Phrases (PP), Noun Phrases (NP), Verb Phrases (VP), Adjective Phrases, and Adverbial Phrases. For other implementations, the word *phrase* may be defined as any proper name (for example "New York City"). Most definitions require that a phrase contain multiple words, although at least one definition permits even a single word to be considered a phrase. Some search engine implementations utilize a lexicon (a pre-canned list) of phrases. The WordNet Lexical Database is a common source of phrases.

When used in conjunction with search engines, the word *concept* generally refers to one of two constructs. The first construct is *concept* as a cluster of related words, similar to a thesaurus, associated with a keyword. In a number of implementations, this cluster is made available to a user - via a Graphic User Interface (GUI) for correction and customization. The user can tailor the cluster of words until the resulting *concept* is most representative of the user's understanding and intent. The second construct is *concept* as a localized semantic net of related words around a keyword. Here, a local or public ontology and taxonomy is consulted to create a semantic net around the keyword. Some implementations of *concept* include images and other non-text elements.

Topics in general practice need to be identified or "detected" from applying a specific set of operations against a body of text. Different methodologies for identification and/or detection of topics have been described in the literature. Use of a topic as input to a search engine therefore usually means that a body of text is input, and a required topic identification or topic detection function is invoked. Depending upon the format and length of the resulting topic, an appropriate relevancy function can then be invoked by the search engine.

Data objects as input to a search engine can take forms including a varying length set of free form sentences, to full length text documents, to meta-data documents such as XML documents. The Object Oriented (OO) paradigm dictates that OO systems accept objects as inputs. Some software function is almost always required to process the input object so that the subsequent relevance function of the search engine can proceed.

Input to the Knowledge Correlation Search Engine differs from current uses because all input modes of the Knowledge Correlation Search Engine must present a minimum of two (2) non-identical terms, phrases, or concepts. "Non-identical" in this usage means lexical or semantic overlap or disjunction is required. The minimum two terms, phrases, or

concepts are referred to as X and Y (or "origin" and "destination"). No input process can result in synonymy, identity, or idempotent X and Y term, phrases or concepts.

As with existing art, text objects and data objects can be accepted (in the Knowledge Correlation Search Engine, as either X or Y) and the topics and/or concepts can be extracted prior to submission to the Knowledge Correlation process. However, unlike most (if not all) existing search engines, the form of the input (term, phrase, concept, or object) is not constrained in the Knowledge Correlation Search Engine. This is possible because the relevancy function (Knowledge Correlation) does not utilize similarity measures to establish relevancy. This characteristic will allow the Knowledge Correlation Search Engine to be seamlessly integrated with many existing IR applications.

Regardless of the forms or methods of input, the purpose of Knowledge Correlation in the Knowledge Correlation Search Engine is to establish document relevancy. Currently, relevancy is established in IR using three general approaches: set-theoretic models which represent documents by sets; algebraic models which represent documents as vectors or matrices; and probabilistic models which use probabilistic theorems to learn document attributes (such as topic). Each model provides a means of determining if one or more documents are similar and thereby, relevant, to a given input. For example, the most basic set-theoretic model uses the standard Boolean approach to relevancy – does an input word appear in the document? If yes, the document is relevant. If no, then the document is not relevant. Algebraic models utilize techniques such as vector space models where documents represented as vectors of terms are compared to the input query represented as a vector of terms. Similarity of the vectors implies relevancy of the documents. For probabilistic models, relevancy is determined by the compared probabilities of input and document.

As described above, the Knowledge Correlation Search Engine establishes relevancy by an entirely different process, using an entirely different criteria than any existing search engine. However, the Knowledge Correlation Search Engine is dependent upon Discovery and Acquisition of "relevant" sources within the corpus (especially if that corpus is the WWW). For this reason, any form of the existing art can be utilized without restriction during the Discovery phase to assist in identifying candidate resources for input to the Knowledge Correlation process.

For all search engines, simply determining relevancy of a given document to a given input is necessary but not sufficient. After all – using the standard Boolean approach to relevancy as an example – for any query against the WWW which contained the word "computer", tens of millions of documents would qualify as relevant. If the user was actually interested only in documents describing a specific application of "computer", such a large result set would prove unusable. As a practical matter, users require that search engines rank their results from most relevant to least relevant. Typically, users prefer to have the relevant documents presented in order of decreasing relevance – with the most relevant result first. Because most relevance functions produce real number values, a natural way to rank any search engine result set is to rank the members of the result set by their respective relevance scores.
Ranked result sets have been the key to marketplace success for search engines. The current dominance of the Google search engine (a product of Google, Inc.) is due to the

PageRank system used in Google that lets (essentially) the popularity of a given document dictate result rank. Popularity in the Google example applies to the number of links and to the preferences of Google users who input any given search term or phrase. These rankings permit Google to optimize searches by returning only those documents with ranks above a certain threshold (called k). Other methods used by web search engines to rank results include "Hubs & Authorities" which counts links into and out of a given web page or document, Markov chains, and random walks.

The Knowledge Correlation Search Engine utilizes a ranking method that is novel because it is a function of the degree to which a given document or resource contributed to the correlation "answer space". That answer space is constructed from data structures called nodes, which in turn are created by decomposition of relevant resources. Even the most naïve ranking function of the Knowledge Correlation Search Engine – which counts the frequency of node occurrence in the answer space – can identify documents that are uniquely or strongly relevant to the original user query. More sophisticated ranking mechanisms can dramatically improve that outcome.

The Knowledge Correlation Search Engine is a new and novel form of search engine which utilizes a computer implemented method to identify at least one resource, referenced by that resource's unique URI (***U**niform **R**esource **I**dentifier)* or referenced by that resource's URL (***U**niform **R**esource **L**ocator),* such resource being significant to any given user question, subject, or topic of a digital information object. For the Knowledge Correlation Search Engine, the user question or subject or topic acts as input. The input is utilized by a software function which attempts to construct or discover logical structures within a collection of data objects, each data object being associated with the resource that contributed the data object, and the constructed or discovered logical structures being strongly associated with the input. That software function is a knowledge correlation function and the logical structure is a form of directed acyclic graph termed a quiver of paths. If such logical structures strongly associated with the input are in fact constructed or discovered, the data object members of such logical structures become an answer space. Using the answer space, another software function is then able to determine with a high degree of confidence which of the resources that contributed to the answer space are the most significant contributors to the answer space, and thereby identify URLs and URIs most significant to the input question, subject or topic. Finally, a software function is used to rank in significance to the input each of the URL and URI referenced resources that contributed data objects to the answer space