Qualitative and Quantitative Methods in Libraries (QQML) 1: 153 –173, 2014

Comparative Evaluation of Three Types of Semantic Distance Metrics – Implications for Use in Semantic Search

Denise A. D. Bedford, Ph.d.[1], and Karen F. Gracy, Ph.d.[2]

[1]Goodyear Professor of Knowledge Management, Information Architecture and Knowledge Management, Kent State University
[2]Assistant Professor, School of Library and Information Science, Kent State University

Abstract. Semantic relatedness is an important measure for search functionality and design in the 21st century. We envision that a 21st century search system should be able to accept as a "query" a sample document or object – and return results which are "like this" or "related." Today, search systems that suggest "related results" do so based on the similarity of values in defined properties or bibliographic fields (e.g., faceted search using metadata values) or on high co-occurrence rates of query terms and full-text indexes. These search systems are commonly referred to as Similarity Search. For search systems to be able to support this capability in the future there must be a reliable mechanism for semantically identifying facets and values in the query document, and for calculating the semantic relatedness or similarity to other documents. The literature is rich with discussions of semantic relatedness and similarity measures. Among the measures discussed, semantic distance appears to hold the greatest promise for this future search capability. Semantic relatedness is a concept that has been treated in philosophy, psychology, artificial intelligence and computational linguistics. This research approaches the concept of semantic distance from the computational linguistics and semantic analysis perspective, e.g., the degree of similarity or relatedness of two lexemes in a lexical resource. Semantic distance provides a more practical and quantitative approach to defining "similarity." In addition, this research expands the definition of a lexical resource to include: full text and text corpus, knowledge organization systems, and metadata structures for documents.

While the literature is rich in discussions of examples and applications of semantic distance measures, a comparative evaluation of those measures against a controlled set of lexical resources is lacking. The purpose of this research is to explore which semantic distance metrics might be most effective, depending on the context and semantic capabilities available.

Received: 19.1.2014 / Accepted: 4.3.2014 ISSN 2241-1925

The research reported by the authors compare the performance of four methods of concept identification and three methods of calculating semantic distance measures in a controlled environment. The four methods of concept identification include (1) human generated metadata; (2) machine guided metadata; (3) statistically guided classification; and (4) deep semantic indexing. The three types of semantic distance metrics include: (1) quantitative translation and interpretation of ANSI/NISO 739.19 standard thesaurus relationships; (2) stochastic co-occurrence of concepts in text corpus; and (3) grammatical relationships. This research builds upon the research that was reported at QQML 2011 (Bedford and Gracy 2011

1. Research Goal and Context

The intent of this research is to identify the most effective semantic representation of a document in the context of similarity search. Similarity search is one of several models of a future semantic search. The context in which this research is undertaken is a future vision of semantic search.

Mills Davis has identified 16 areas of that semantic landscape that are essential to the development of this foundation. One of these is semantic search or what Davis labels "From Search to Knowing." (Davis 2008) (Davis 2011). By semantic search we generally understand any form of search that leverages semantic capabilities and features. Semantic search is not a single application, nor is it a single perspective. Grimes (Grimes 2010) described eleven models of semantic search. Seven of these models apply semantics to query (Table 1). These models of semantic search expand the context and meaning of search terms. Four of these models leverage semantics to augment search results (Table 2). These models of semantic search organize and highlight results in a meaningful and contextual way.

This research focuses on one of these eleven models – full-text similarity search. According to Grimes and the published literature, this model currently makes little if any use of semantics. What is full-text similarity search? And, why is it important? Similarity search forms have two common elements – they identify a common set of features among a population of objects and they leverage a definition of "relatedness" to find similar objects. Some forms of similarity search begin with a selection or example offered by the searcher. Others begin with a population or a defined set of objects.

Similarity searching is important because it allows a search system to build a rich representation of what the user prefers or is looking for – based on an example provided by the user. Similarity search based on a good example has the potential to be more effective than simple keyword searching or controlled parametric searching. It has the potential to improve the search result without increasing the burden on the user. We believe, though, that this potential can only be realized where semantics are used to provide a robust representation of the underlying content of the example. Such a robust representation may be transformed into a multifaceted query.

Table 1.
Grimmes Examples of Semantic Search Focused on Query

Type of Semantic Search Application	Description
Reference Advice	Results include materials that may provide further information about the terms used in the query.
Search on Semantic or Syntactic Annotations	Query processing approach which either implicitly or explicitly semantically tags query terms to improve the context in which the query terms are searched.
Concept Search	Search which expands the query terms to include semantically related concepts. The expansion can be done explicitly or implicitly.
Ontology-Based Search	Search system assigns semantic meaning to the query terms and associates the terms with other terms that may be related in other contexts.
Semantic Web Search	Searching complex relationships that will be available in the future web of data.
Faceted Search	Otherwise known as parametric or fielded search. Search is enabled for more than one dimension – multiple facets or parameters that enable a search to more precisely define their search needs.
Natural Language Search	Search query is semantically processed and tagged for more effective matching against other objects in the search system index.

Table 2.
Grimmes Examples of Semantic Search Focused on Search Results

Full-Text Similarity Search	Search uses a submitted block of text or a full document to identify other results which may be similar. Similarity is generally determined based on statistical or vector-space similarity measures. There is typically no "semantic meaning" associated with the similarity ratings,
Related Searches and Queries	Search suggestions that highlight objects that are similar in some way to the query terms. This can either be explicit or implicit suggestion.
Semantically Annotated Results	Search results have highlights for terms in the documents that are semantically-related to the search query.
Clustered Search	Search results are statistically clustered into categories to help the searcher more effectively navigate the results sets.

2. Current State of Full-Text Similarity Search

Today there are three forms of similarity search. Each form has a method for extracting features and determining relatedness. The first form includes recommendation engines where features or characteristics of objects are based primarily on descriptions and interpretations contributed by people. Recommendation systems have been design for books (i.e., Amazon.com), music (i.e., Pandora, Spotify, Last.fm) and people (i.e., Match.com). A multifaceted query constructed from the feature set is used to identify related objects. Relatedness is more often than not a simple metric of match/no match.

The second form of similarity search includes vector-based applications (Giunchiglia et al 2004) (Li et all 2011) (Otlacan and Otlacan 2006) (Roddick et al 2003) (Tsang and Stevenson 2008) (Weber and Schek 1998) (Zezula et all 2006). In this context, characteristics or features used for matching are deduced from the population of objects in the defined space or the data set. Several statistical methods may be used to define the relatedness or closeness of objects within the space, including K-nearest neighbor, nearest neighbor search, proximity search, approximate nearest neighbors (ANN), range queries, maximal intersection queries, post-office problem, partial match, best match file searching, best match retrieval, and sequence nearest neighbors (SNN).

As Grimmes and others suggest, though, there is little or no use of semantics either in the extraction of features or the determination of relatedness in this form of similarity search.

The third form of similarity search is used in file compression and storage contexts. The goal in this context is to reduce the amount of space required to store information by eliminating similar redundant content. This approach involves a deep full object semantic analysis. However, the analysis is discarded the compressible text has been identified. Relatedness of objects is important only as an intermediate step to compression.

3. Research Questions

The essence of the research reported in this paper is the construction of a robust semantic representation of the sample document. By semantic representation we mean the implicit semantic networks comprised of concepts (e.g., nodes), the connectedness of those concepts through semantic links, and measures of semantic distance associated with the links (Figure 2). A robust semantic representation of a document may be transformed into a complex query to identify and match other documents. The research poses and investigates three questions, including:

> Question 1. What are the key elements of a semantic representation of a document?
>
> Question 2. How should concepts be represented to support similarity searching?
>
> Question 3. How should relationships among concepts in a document be represented to support similarity searching?

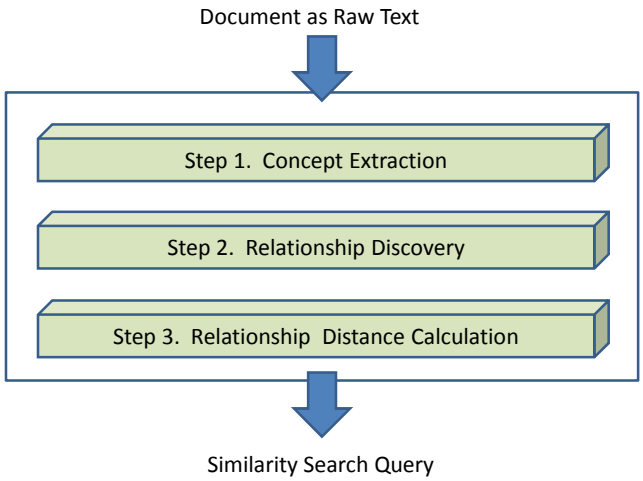
Research Data

This research builds upon work that was reported at QQML 2011 (Bedford and Gracy 2011). The original data set created in 2011 serves as the source data set for 2012. The research data includes 247 full text documents derived from the World Bank's publicly available Documents and Reports database. The document set is accessible through the Documents and Reports database using machine generated metadata and the World Bank Thesaurus 2007 edition.

4. Research Methodology

A three step methodology was designed to investigate the research questions (Figure 1). Step 1 involved development of a reference model to use in creating a semantic representation of each document in the data set. Step 2 involved feature extraction to identify the nodes in each reference model. Four methods were used to generate feature extraction. Step 3 involved the identification and comparison of three methods for identifying relationships among concepts (e.g., edges and nodes), and for assigning distance values to each relationship (e.g., edge). As a final consideration, we explore the transformation of the fully elaborated semantic profile into a complex query. The models and methods are discussed in detail below.

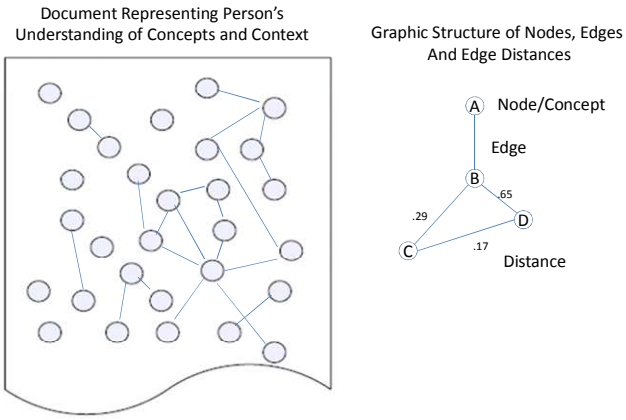
Figure 1.
Three Step Research Methodology



Question 1: What are the key elements of a semantic representation of a document?

We suggest that similarity searching is most effectively achieved when the query takes the form of a deep semantic representation of a document. By a semantic representation we mean a semantic network representation of the concepts and ideas expressed in the document, and the relationships among concepts that describe the context.

Figure 2.
Semantic Representation of Single Document



We suggest that a document can be represented as a semantic network structure, where concepts are network nodes, where relationships among the concepts are network edges or links, and where the value of relationships are distance values (Figure 2). The semantic representation of a document is an aggregation of all the nodes, the edges or linkages among the nodes and the value of the linkages (Table 1).

We used standard network metrics to evaluate the semantic richness of a document. We suggest that the metric for nodes is a simple count. This count is equivalent to the number of concepts in the query. The metric for edges is determined based on the number of links in relation to nodes. Nodes with a greater number of links or edges (e.g., *network centrality)* would be given greater weight in similarity search. The number of edges would be used to assign weights to concepts in a similarity search. Finally, the value of the edge between two nodes - what is generally described as semantic distance – tells us something about the strength of the relationship between two concepts.

We expect that different methods of representing documents will produce different numbers of nodes, of links, and different link values.

Table 1
Logical Representation of the Semantic Network of a Document

Semantic Network Factor	Metric	Pertinence to Similarity Search
Concept Nodes	Number of Nodes	Key concepts to be searched
Edges or Links	Number of Unique Edges	Weights assigned to a concept
Edge Value	Qualitative interpretation Quantitative definition	Value of the weights – pre-defined relevance between concepts
Network Centrality	Ratio of Nodes to Edges	General semantic density of the document for evaluation purposes
Edge Value Variation	Variation of edge values in the overall network	Validation of meaningful distance values

Variations may influence how well similarity search may perform. Where there are fewer nodes, the similarity search will have fewer matching points. Where relationships are sparse, the understanding of the context will be poorer. Sparse network models may be poor or high risk candidates for similarity search. Where semantic distance varies, the relevance or relatedness of two concepts would vary.

Question 2. How should concepts be represented to support similarity searching?

Similarity searching with a sample document may provide rich description of the information the searcher is looking for. We evaluate four methods of identifying document concepts to understand how well they represent the content of the document. Each of the four methods may be used to generate a surrogate of the document for similarity searching. These four methods were chosen because they are in common use today. The four methods are described in Table 2.

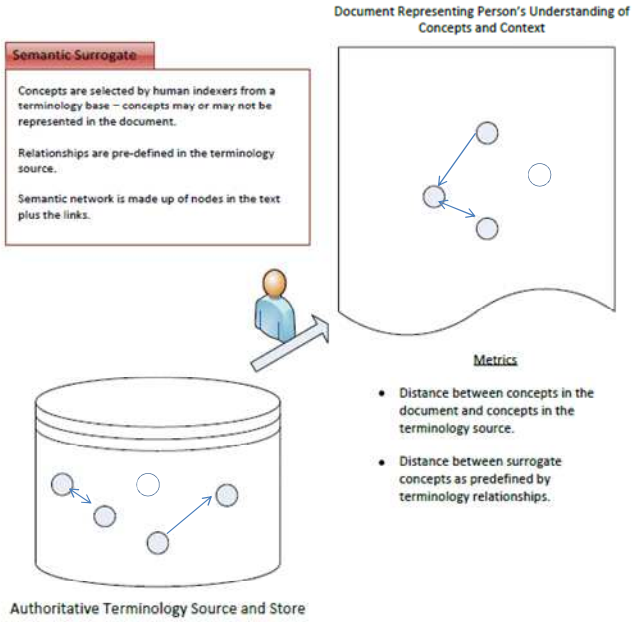
Table 2
Four Methods for Identifying Document Concepts

	Method A	Method B	Method C	Method D
Knowledge Organization System Used	Library of Congress Subject Headings	World Bank Thesaurus	N/A	N/A
Concept Selection Method	Human Indexer's Decision	Automated Indexing	Statistical Clustering	Raw Concept Extraction

Description of Method A: Thesaurus as Concept Source and Human Indexing

Method A (Figure 3) represents human generated metadata as represented in the metadata present in MARC records in the OCLC database. The knowledge organization system that guided the selection of concepts was the Library of Congress Subject Headings authority. While this representation is readily available for many documents, it does not appear to provide a rich representation of the concepts in the document. The average number of concepts identified per document using this Method was 3.36. The maximum number was 11, and minimum was 0.

Figure 3
Method A. Subject Headings as KOS and Human Indexing

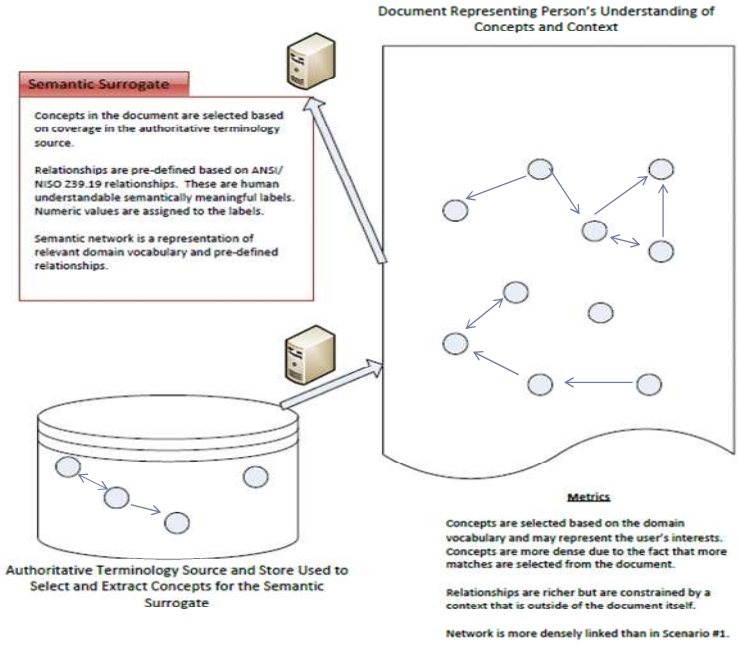


Description of Method B: Thesaurus as Concept Source and Semantic Technology Indexing

Method B (Figure 4) represents concepts extracted from metadata records published in the World Bank's Document and Records database. The knowledge organization system that guided the selection of concepts was the World Bank Thesaurus 2007 Edition. The method of selection of concepts was the SAS Content Categorization Suite, which supported automated indexing guided by embedded knowledge organization systems. All concepts selected were explicitly present in the document. No intermediate interpretation of concepts was used. The average number of concepts identified per document

using this Method was 115.77. The minimum number of concepts generated for short documents was 3, and the maximum was 245. The Method does produce a rich set of concepts representative of the document content.

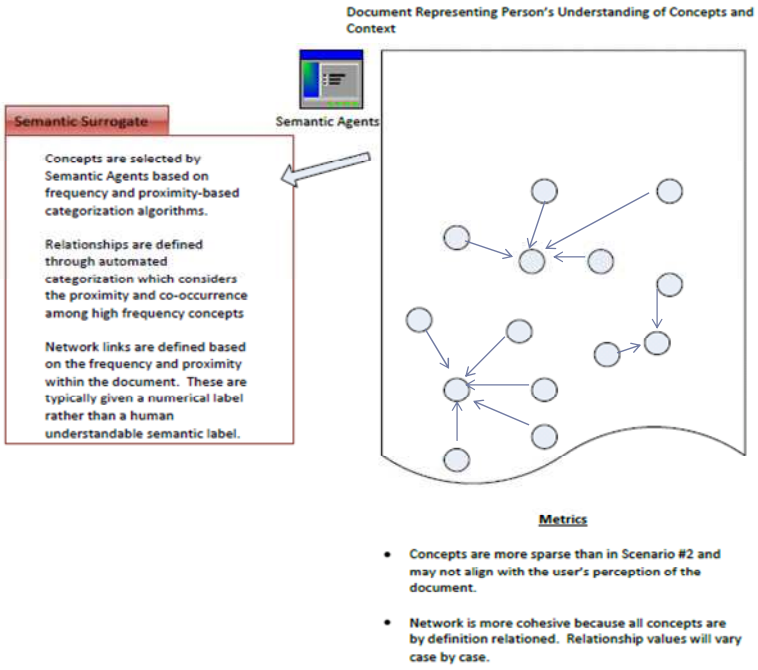
Figure 4
Method B. Thesaurus as KOS and Automated Indexing



Description of Method C: Concept Discovery and Organization Using Statistical Clustering

Method C (Figure 5) represents concepts selected directly from the document. No knowledge organization system that guided the selection of concepts. The method of selection was a computer algorithm which leveraged frequency and co-occurrence of concepts. ClearForest was the semantic technology that supported this method. All concepts selected were explicitly present in the document. No intermediate interpretation of concepts was used. The average number of concepts identified per document using this Method was 16.41. The minimum number of concepts generated for short documents was 2, and the maximum was 63. Method C is clearly more robust than Method A, but does not produce as rich a representation as Method B.

Figure 5
Method C. Concept Discovery and Organization Using Statistical Clustering



Description of Method D: Deep Semantic Description of Content

Method C (Figure 6) represents the full set of concepts represented in the document. This is the equivalent of the number of words in the document. This method is used to create semantic signatures for storage compression purposes. While it is very robust in terms of identifying exactly equivalent or closely equivalent documents (i.e., versions or editions), it is not a good candidate for similarity search. The number of concepts extracted would be far too dense to use for similarity search. The query could be comprised of as many as 5,000 concepts. This would be inefficient from both a query and a matching perspective.

The research results for Question 2 would suggest that Method B would be the best candidate for producing concepts for similarity search. Method C might also be acceptable, though there is a risk associated with the fact that the user perspective is not represented in an embedded knowledge organization system. Method A representing a traditional metadata representation is not sufficient for similarity search. Method D representing a deep semantic signature approach would overwhelm a similarity search.

Figure 6.
Method 4. Deep NLP Representation

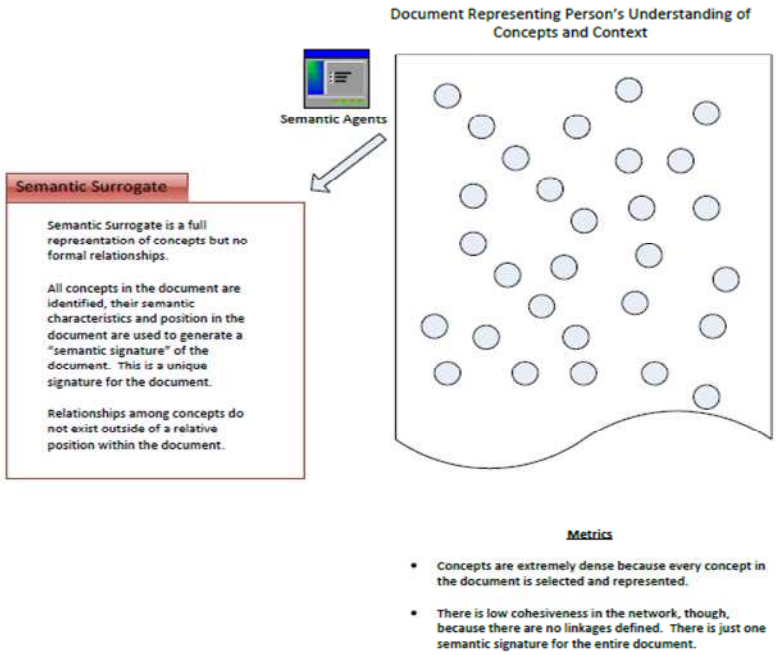


Table 3
Results Summary for Concept Discovery

	Method A	**Method B**	**Method C**	**Method D**
Average Number of Concepts	3.36	115.77	16.4	Ca. 7,000
% Document Concepts Represented	.04%	1.6%	.22%	100%
Concept Representation	Extremely Sparse Nodes, Many not explicitly	More Nodes – All Explicitly Available in the	Small Number of Nodes – All Explicitly Available in	Overwhelming Number of Nodes – No Selectivity or Distinction

	Method A	**Method B**	**Method C**	**Method D**
	present in the document	Document	the Document	
Evaluation for Similarity Search	Poor Candidate – Insufficient Content for Matching	Strong Candidate – Manageable Number of Nodes for Search and Matching	Acceptable but Suboptimal – Manageable Number of Nodes for Search and Matching. No user representation in concept selection.	Poor Candidate – Unmanageable for Search, and Inefficient for Matching

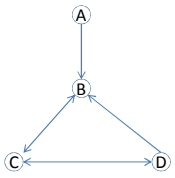
Question 3. How should relationships among concepts in a document be represented to support similarity searching?

The second important aspect of semantic representation is the relationships among concepts. Relationships, or network edges, describe the context in which the concepts are used. Relationships are important for understanding why the sample document is a good query candidate for the searcher. It is important to have a rich set of concepts as a baseline. The context in which those concepts are discussed is what differentiates a similarity search from a keyword or parametric search. Figure 7 below tell us something about the context of a document through the structure of the links.

Knowing what relationships exist among concepts is a minimum requirement for similarity search. Even more important, though, is the nature and strength of the relationships between concepts. Similarity search across sources requires at least a framework for understanding. How is the distance between nodes (i.e., the value of the edge) calculated? How would it be interpreted across documents?

From the perspective of similarity searching, relationships or the number of edges that are assigned to nodes provides an indication of node weight (Figure 8). This translates to query concept weighting. The distance value of an edge between two concepts tells us something about the strength of the relatedness. Intuitively, concepts that are highly related will have a high value associated to the link (e.g., a short distance conveys high relatedness and high value).

Figure 7.
Network of Concepts (Nodes) and Relationships (Edges)



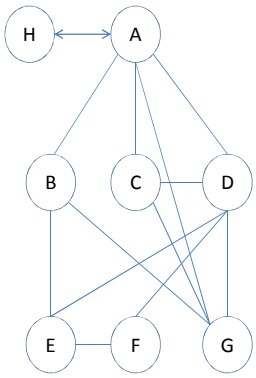
Directed Graph with 4 Nodes

A = Biodiversity
B = Marine Biodiversity
C = Marine Habitat Conservation
D = Cruise Ship Effluent

A to B = Hierarchical Relationship
B to C = Associative Relationship
C to D = Associative Relationship
D to B = Associative Relationship

Figure 8.
Common Network Metrics

Example of a Network (Nodes and Edges)



Network Connectivity =
Total Number of Edges/Links
(This Example = 11)

Degree Centrality =
Number of Unique Edges
for a Single Node
(This Example: A = 4)

Edge Distance/Weight =
Aggregate Value of Edges
(Value of each edge is
determined, then all values
are summed)

Semantic Edge and Edge Distance Evaluation – Method 1

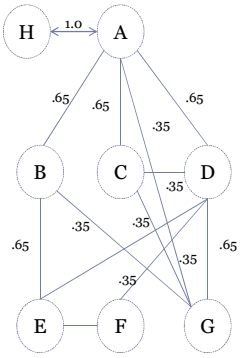
We evaluated two methods which are in wide use today. The first method leverages pre-defined, human understandable relationships such as those described in the ANSI/NISO Z39.19 Standard (NISO 2005). These relationships are expressed as "classes" or "types" of relationships – and labeled as equivalent, hierarchical, and associative relationships. These relationships are not mapped to semantics (e.g., verb phrases). That interpretation is left to humans. These relationships are generally not applied to single documents. Rather, they are represented as domain-specific knowledge organization systems which are consulted by indexers. Indexers select and apply concepts, but not the relationships. Most search systems that use these relationships for query expansion assign a single numerical value to the relationships in the knowledge organization system. Common weights are: (1) 1.00 for equivalence relationships; (2) .65 for hierarchical relationships; and (3) .35 for associative relationships. Figure 9 describes the semantic context of a document in terms of its expressed relationships. It illustrates the degree of centrality for concepts and the individual edge values.

In order to discover the edge values for the documents in the data set, a document-specific thesaurus was created for each document. Some thesauri were very sparse – containing only three concepts – but others were substantive. Each thesaurus was created following strict interpretation of the ANSI/NISO Z39.19 Standards for establishing thesaurus relationships. Each thesaurus was created in the MultiTes Thesaurus Management software. MultiTes automatically generates statistics for the number and type of relationships established. Statistics were collected for each thesaurus. Consistent with common practice, equivalence values were assigned a value of 1.00. Hierarchical relationships were assigned a value of .65, and associative relationships were assigned a value of .35.

Semantic Edge and Edge Distance Evaluation – Method 2

The second method identifies relationships based on their proximity within a document. The relationships are assigned simple numerical values that convey something about the "nearness" of the concepts. These values don't convey meaning that a human can interpret. But, they can tell us that the related concepts are important for understanding the context of the document. This method is used by clustering engines, automated topic mapping systems, and some statistical categorization engines. Each application has its own embedded algorithm for defining distance values. Figure 10 illustrates extreme concept centrality of this method. It also illustrates the variant weights assigned to relationships among concepts.

Figure 9.
Weighted ANSI/NISO Z39.19
Relationships



Network Connectivity = 10 Edges/Links

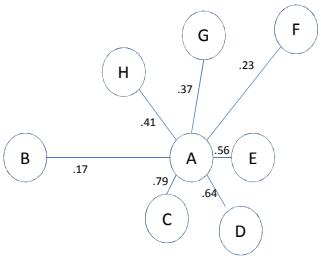
Degree Centrality = No. of Unique Edges for a single node

Degree Centrality A, G = 4 Edges
Degree Centrality D = 5 Edges
Degree Centrality B, C, E = 3 Edges

Edge Distance/Weight = Aggregate Value of Edges

Distance A - B = .65
Distance A - C = .65
Distance A - E = 1.3
Distance B - G = .35
Distance A - G = .35
Distance A – H = 1.00

Figure 10.
Relationships Discovered and Valued Using Automated Categorization Algorithm



Network Connectivity = Total Number of Edges/Links

Network Connectivity = 7

Degree Centrality = No. of Unique Edges for a single node

Degree Centrality A = 7
Degree Centrality B, C, D, E, F, G, H = 1

Edge Distance/Weight = Aggregate Value of Edges

Distance A - B = ..17
Distance A - C = .79
Distance A - D = .64
Distance A – E = .56
Distance A - G = .37
Distance A – H = .41

Based on the research results for Question 2, Method A and Method D were not considered to be worthy of further evaluation. In fact, the concepts generated by Method A were insufficient to establish relationships. The concepts generated by Method D were too numerous to generate meaningful relationships. The evaluation of Question 3, therefore, focused on Methods B and C.

Question 3: Evaluation of Results

We evaluated the research data generated for both methods in terms of (1) total number of links discovered within the document (Network Connectivity); (2) average aggregate value of the edges; and (3) the average value of link (semantic distance). It is clear from the results (Table 4) that Method B generates a much richer set of relationships and thus a much richer semantic network structure. The average number of links per document – illustrating context – is high at 78.87. This is five times greater than the value generated by Method C, the automated categorization algorithm. Similarly, the average aggregate value of links for a document generated by Method B– the sum of all links - is also high at 31.95. This is ten times greater than that the aggregate value generated by Method C.

Table 4
Comparison of Two Methods for Establishing and Valuing Relationships

	Method B	Method C
Method of Defining Relationships	ANSI/NISO Z39.19 Relationship	Document Position and Statistical Co-occurrence
Values Assigned to Relationships	Pre-defined and static values for individual relationships (Figure x)	Each relationship has a unique numerical value (Figure)
Average Number of Links Per Document (Network Connectivity)	78.87	15.41
Aggregate Value of Links Per Document (Ave. for all Documents)	31.95	3.01.
Value of Link – Semantic Distance (Ave. for all Documents)	.381	.21

The unexpected result, though, pertains to the focus of this research. The variability of is the relative consistency of the average value of links across the two methods. The pre-defined static values generated by the individual document thesauri had an average value of .381, with a high of .7 and a low of 0 (Table 4). Method C produced an average value of .21, with a high of .49 and a low of .08. What variability we see in the results is generated by the distinction

between simple related concepts which have a longer average semantic distance and the equivalence and hierarchically related concepts which have much shorter semantic distances. However, it is also clear semantic relationships created in both Method A and B are more likely to be associative in nature. This raises a question about the relative importance of distinguishing between these two methods.

5. Research Results

What do these results suggest for enhancing the semantics of similarity search? We believe the results suggest that creating a rich semantic representation of concepts is important. A network representation of a document which is rich in concept nodes may provide a rich foundation for similarity search. The most promising method for achieving this may be automated indexing systems with embedded knowledge organization systems. The other three methods evaluated clearly produced inferior results. Where a rich representation of concepts has been created, either Method A or Method B could be used to generate relationships. The average semantic distance between concepts, the value of network edges, did not favor one method over the other.

Discussion of Results

The research suggests that it is possible to represent a document as a semantic network, where concepts take the role of nodes, relationships represent edges, and values are assigned to the edges. Such a semantic network structure can also be transformed into a rich query for the purpose of similarity search. For use in similarity search, concepts are the most important component of the structure. If the query is not richly representative of the document, similarity search will be no better than keyword or parametric search.

The research suggests that there is more than one way to effectively discover relationships among concepts. The richness of the relationing or the connectedness of the semantic network structure depends largely on the richness of the concept base.

The result which was unexpected pertains to the values assigned to relationships. There appears to be a logical consistency in the way that values are assigned. In this research, only two methods were tested. However, these two methods represent the approaches that are commonly used in the field today. Acknowledging the close alignment of these two methods, the next step in this research would be the development of a reference framework for semantic relationships. Such a framework could serve as a translation tool for similarity searching in a semantic grid.

References

National Institute of Standards Organizations (2005). **ANSI/NISO Z39.19 - Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies**. Bethesda Maryland.

Bedford, D. A. D. and Gracy, K. F. (2011). Leveraging Semantic Analysis Technologies to Increase Effectiveness and Efficiency of Access to Information. *QQML 2011*

Billingham, P. (2007). Searching for the Answer 2.0. *Legal Information Management* Vol. 7, 258-262.

Budanitsky, A. and Hirst, G. (1999). Semantic distance in WordNet: an experimental application-oriented evaluation of five measures. (Accessed online on April 30, 2012 at).

Carullo, M., Binaghi, E. and Gallo, I. (2009) An online document clustering technique for short web contents. *Pattern Recognition Letters* Vol. 30, 870-876.

Cooper, M. (2000). Semantic distance measures. *Computational Intelligence* Vol. 16, No. 1, 79-94.

Cormack, G. (2010). Keyword Searches Disappoint. *Computerworld* December 6, 2010

Daoud, M., Tamine, L. and Boughanem, M. (2011). A personalized search using a semantic distance measure in a graph-based ranking model. *Journal of Information Science* Vol. 37, No. 6. 614-636.

Davis, M. (2008). *Semantic Wave 2008 Report: Industry Roadmap to Web 3.0 & Multibillion Dollar Market Opportunities*. October 2008. http://www.eurolibnet.eu/files/Repository/2009050/165103_SemanticWaveReport2008.pdf (Accessed Nov. 7, 2011)

Davis, M. (2011) *Web 3.0 Manifesto: How Semantic Technologies in Products and Services Will Drive Breakthroughs in Capability, User Experience, Performance, and Life Cycle Value.* http://project10x.com/about2.php (Accessed Nov. 11, 2011)

Dietze, H. and Schroeder, M. (2008). Semantic search engine for the Life Science Web. Presentation delivered to e-Science Institute, Edinburgh, Scotland. November 28, 2008. (Accessed online on April 30, 2012 at:)

Dong, X., Halevy, A., Madhavan, J., Nemes, E., and Zhang, J. (2004). Similarity Search for Web Services. *Proceedings of the 30th VLDB Conference*, Toronto, Canada, 2004.

Giunchiglia, F., Shvaiko, P. and Yatskevich, M. (2004). S-Match: An algorithm and an implementation of semantic matching. Technical Report #DIT-04-015. University of Trento, Department of Information and Communication Technology.

Grimmes, S. (2010). Breakthrough Analysis: Two + Nine Types of Semantic Search. *InformationWeek* January 21, 2010.

Guha, R., McCool, B. and Miller, E. (2003). Semantic Search. *WWW2003* May 20-24, 2003 Budapest, Hungary.

Haveliwala, T. H., Gionis, A., Klein, D. And Indyk, P. (2002). Evaluating strategies for similarity search on the web. *WWW2002* May 7-11, 2002. Honolulu, Hawaii, 2002.

Imielinski, T. and Signorini, A. (2009). If you ask nicely, I will answer: semantic search and today's search engines. *2009 IEEE International Conference on Semantic Computing*

Li, Y., Chu, V., Blohm, S., Zhu, H., and Ho. H. (2011). Facilitating Pattern Discovery for Relation Extraction with Semantic-Signature-based Clustering. *CIKM'11,* October 24–28, 2011, Glasgow, Scotland, UK.

Otlacan, E. and Otlacan, R.-P. (2006). Informational topology and globalization process. *Kybernetes* Vol. 35, No. 7/8, 1203-1209.

Rada, R. H., Mili, E. Bicknell and M. Billettner (1989). Development and application of a metric ton semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*. Vol. 19, No. 1, 17-30.

Roddick, J. F., Hornsby, K. and de Vries, D. (2003). A Unifying semantic distance model for determining the similarity of attribute values. *Twenty-Sixth Australasian Computer Science Conference (ACSC2003)*, Adelaide, Australia.

Shvaiko, P. and Euzenat, J. (2004). A survey of schema-based matching approaches. Technical Report#DIT—04-087. Department of Information and Communication Technology, University of Trento.

Spree, U., Feiszt, N., Luhr, A., Piesztal, B. Schroeder, N., and Wollschlager, P. (2011). Semantic Search – State-of-the-Art Uberblick zu semantischen Suchlosungen im WWW. **Handbuch Internet-Suchmaschinen 2**.

Sussna, M. (1993) Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKKM-93)* Arlington VA 67-74.

Tsang, V., and Stevenson, S. (2008). A graph-theoretic framework for semantic distance. *Computational Linguistics* Vol. 36, No. 1, 31-69.

Weber, R. and Schek, H.-J. (1998). A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. *Proceedings of the 24th VLDB Conference*. New York, USA, 1998.

Zezula, P., Amato, G., Dohnal, V. And Batko, M. (2006**). Similarity Search - The Metric Space Approach.** Advances in Database Systems, Vol. 32. Springer, 2006.