Preface

APPROXIMATE METHODS IN SEMANTICS FOR THE INTERNET: AN INTRODUCTORY MAP TO QUANTITATIVE SEMANTICS

This preface offers a conceptual framework for putting in context the extensive work that has been done in areas related to semantic similarity measuring, approximate semantic methods, document clustering and grouping, and many others that we call generically "Quantitative Semantics". We present some of the most relevant works in these areas, classifying them so that a kind of map emerges, allowing a reader to understand the bigger picture of this tremendously relevant area in many text-processing tasks.

As part of this framework, we provide a brief explanation for the works herein exposed, giving to the reader a scaffolding structure where the works in this book can be conceptually attached.

INTRODUCTION

The Internet has been acknowledged as one of the recent technological revolutions due to its large impact on the whole society. Nevertheless, precisely due to its impact, limitations of the Internet have become apparent; in particular, its inability to take into account in an automatic way the meaning of online documents.

There are initiatives aiming to enhance the Internet, by introducing new languages, which are capable of expressing in a self-contained way the meaning of online documents. They have been called the "Semantic Web" (SW) (Antoniou, Van Harmelen 2004), and have delivered an impressive set of standards, like RDF, OWL and more, which are the base for writing "ontologies", as a way to encapsulate semantic units.

Nevertheless, several practical problems have hindered the Semantic Web from becoming mainstream. Today, some 10 year after been proposed in a manifesto, the SW-enabled web pages do not represent even a small fraction of the whole online available content; they are roughly a quarter of one percent of the whole web. Even duplicating its volume each year, SW would barely reach 1% of the web in two years, which is an overly optimistic estimation.

Many reasons have been argued for SW limited growth (Marshall, Shipman 2003), like SW technological sophistication requirements on users, the quantity of work required to rebuild the web on SW terms, or even (circularly) blaming the SW marginality for its little usability –which makes it marginal.

This book is inspired by the desire to find and extract useful information in Web-based documents, which are a wealth of readily-available information. Instead of waiting for the SW to become mainstream, we can analyze and process the actual content of today's Inter-net. At the least, extracting meaning from the actual Internet could complement and pave the way to SW future developments.

We wish to access the meaning residing in the documents, not just to find out if they contain certain key words or are sitting in highly popular web sites. So we try to answer the question of how to model the semantics of a document, and how to extract (useful) knowledge from one of them. Or, put in other way, how to model the semantic of words, how to extract (useful) knowledge from their appearance in actual documents?

One of the backbone concepts we will use in this chapter for presenting and comparing the different works included in this book, is the notion of "semantic similarity", which tries to give specific measures for the idea that two or more words or texts could refer to the same or similar topic. This concept is introduced in the next section, and then, in the following subsections we delve into three forms of semantic similarity, for documents, words, and concepts. After that, we will discuss the use of semantic similarity in information retrieval tasks such as search.

While discussing the above, we will present specific works for each of the items, using a label of the form "W.x" for a number x, -"W" standing for "work"- including of course the chapters of this book.

So, in this introductory chapter, our goal is to place the chapters of the book in the map of current literature of semantic similarity, not being exhaustive, but citing some relevant works for each topic.

SEMANTIC SIMILARITY

The semantic similarity between documents and terms has been studied in the fields of at least information retrieval and machine learning, and there is a whole body of literature on measuring the semantic similarity (Resnik, 1995; Landauer, Dumais 1997; Lin, 1998; Hatzivassiloglou et al. 1999; Turney, 2001; Maguitman et al. 2005; Ramirez, Brena, 2006; Sahami et al. 2006; Bollegala, 2007). These methods can be roughly classified into two major categories: knowledge-based approaches (§1.1.1) and corpus-based approaches (§1.1.2).

Before answering the questions asked in the preceding paragraphs, we must query again: What for? What is the goal of these models, of the intended extraction? Or, in equivalent form, how will I know if the constructed models are adequate, are appropriate? Do the models match the modeled documents, regarding the information in them contained? How will I know if the correct conclusions were obtained from the accessed documents (or from those reported as "relevant")? We can recognize four important goals:

- Finding similarity among documents (§1.1);
- Finding similarity among words or symbolic values (§1.2);
- Finding similarity among *concepts* (not words) (§1.3);

Semantic Similarity among Documents

Automatic methods to compute similarity between texts have applications in many information retrieval related areas, including natural language processing and image retrieval from the Web, where the text surrounding the image can be automatically augmented to convey a better sense of the topic of the image.

Here we place the efforts to group documents in collections of "semantically similar" documents. This grouping or clustering can be done by either

- Attaching them a specific label or tag –or using a predefined taxonomy or collection of such tags (§1.1.1); or by
- Grouping or clustering the documents, but without use of predefined labels or taxonomies (§1.1.2).

These two alternatives correspond to the classification of semantic similarity into two main approaches that we mentioned before: Knowledge-based methods and corpus-based methods.

Knowledge-Based Methods

Knowledge-based approaches are based on the use of predefined directories or ontologies to identify semantic relations. Measures of semantic similarity between documents take as a starting point the structure of a directory or document classification scheme where the entities have been previously sorted. Examples of such classifications include WordNet (See W.19) for the case of terms, and Dmoz, the Wikipedia's List of Academic Disciplines or the Yahoo directory, for the case of documents.

These methods attach to the documents a specific label or tag: the theme or topic they cover, the "name" of the collection. When the methods use human-defined categories for topics (Among them, W.3, W.4, W.6, W.7, W.8) they become prone to problems of disagreement, difficulty to update, etc.

Keeping up-to-date the ontologies used by the knowledge-based approaches is expensive. For example, the semantic similarity between terms changes across different contexts. Take for instance the term java, which is frequently associated with the java programming language among computer scientist. However, this sense of java is not the only one possible as the term may be referring to the java coffee, the java island or the java Russian cigarettes, among other possibilities. New words are constantly being created as well as new senses are assigned to existing words. As a consequence, the use of knowledge-based approaches has disadvantages because they require that the ontologies be manually maintained. Alternatively, methods that disambiguate words into meanings (See W.2) can be used.

W.1. Some organizations have tried to organize pages in a predefined classification, and have manually built large directories of topics (e.g. Dmoz or the Yahoo! directory). But given the huge size and the dynamic nature of the Web, keeping track of pages and their topic manually is a daunting task. There is also the problem of agreeing on a standard classification, and this has proved to be a formidable problem, as different individuals and organizations tend to classify things differently.

W.2. Automatic methods to identify the topic of a piece of text. Clasitex (Guzman 1998) finds the main topics that a document talks about. When visiting in a left-to-right one pass each word of the document, counters are incremented. But Clasitex does not count the words, but the *concepts* to which they refer. Since words are ambiguous, a word usually increments several counters (one for each concept the word denotes). Thus, without really disambiguating, concept counting is achieved (instead of word counting as in Support Vectors (See W.10)).

W.3. Automatic methods to identify the topic of a piece of text have also been used in text summarization (Erkan and Radev 2004),

W.4. Automatic methods to identify the topic of a piece of text have also been used in text categorization (Ko et al. 2004),

W.5. A method that automatically assigns labels to clusters of retrieved documents (using keywords and perhaps several search engines) is described in (Bordogna *et al*, *this book*). It accomplishes this task by using labels at different levels of granularity. Implicit topics are revealed by associating labels with the retrieved items, the clusters, and the retrieved lists. Then, some manipulation operators are applied to each pair of retrieved lists, clusters, and single items, to reveal their implicit relationships. The outcome is represented as multi-granular graphs, at three granularity levels. The algorithm uses Latent Semantic Analysis (see W.11).

W.6. Automatic methods to identify the topic of a piece of text have also been used in word sense disambiguation (Manning, Schutze 1999).

W.7. Automatic methods for identifying the topic of a text have been used in evaluation of text coherence (Lapata and Barzilay 2005)

W.8. Automatic methods to identify the topic of a piece of text have also been used in automatic translation (Liu and Zong 2004),

W.9. If we retrieve documents using some form of contextual search, how do we evaluate the semantic similarity of the retrieved documents with respect to the contextual search specified? How well was the retrieval achieved? (Maguitman, Lorenzetti and Cecchini, *this book*) propose a way to evaluate contextual search using a framework that takes advantage of semantic similarity data. For this purpose, they first review the graph-based measures of semantic similarity. They use an incremental algorithm that uses topic descriptors and discriminators for evaluating the described semantic similarity (Lorenzetti & Maguitman, 2009).

One common method for measuring relatedness of two classified entities is to measure the distance between them when traveling in the hierarchy from one to the other (Budanitsky, Hirst 2006). For example, let us consider the Wikipedia hierarchy of Academic Disciplines. Consider the “Machine learning” and “Functional programming” articles; by examining the list of disciplines, a path for each one can be extracted (in our case, to facilitate distance measuring when two pages fall into completely different disciplines, a root element called Areas has been included). They both belong to the area of Computer Science; so, in order to get to Functional Programming one would have to go 1) from Machine Learning to Cognitive Science, 2) from Cognitive Science to Artificial Intelligence, 3) from Artificial Intelligence to Computer Science, 4) from Computer Science to Programming Languages, and finally 5) from Programming languages to Functional Programming. The number of hops was five. It is possible to refine this kind of metrics by weighting the distance according to the depth in the hierarchy, in order to give more importance to branching occurring high in the hierarchy. It has also been proposed to define the “familial distance” d_f (s, d) from a source document *s* to another document *d* in a class hierarchy be the distance from *s*’s class to the most specific class dominating both *s* and *d* (Haveliwala et al., 2002); the problem with this last metric is that it is not symmetric, that is, d_f (s, d) is not always the same as d_f (d, s). This problem has been studied in the context of words (symbolic values) placed in a hierarchy (see W.25).

However, measuring semantic proximity based on hierarchies has several drawbacks. First, it is necessary to incorporate to the hierarchy all and every one of the documents to be considered. But the coverage of any extensive directory is known to be rather small; less than 4 million pages are in any directory, out from billions indexed by Google alone. Next, keeping hierarchies up-to-date manually is difficult, as the growing pace of the Internet is of at least 25 millions of websites per year. Automated approaches have been proposed for automatic incorporation of new web pages (Pierrakos et al., 2003), using some of the methods we discuss later.

Corpus-Based Methods

These methods work without a hierarchy of terms, of without specifying the topic that groups the documents in a collection. That is, these documents are semantically similar since they talk about certain topic, but the topic is not specified, so the similarity is more of a "statistical" nature, as the documents in a class contain the same proportion of words, or the same groups of words.

W.10. Support vectors (Joachims 2002). Words of a given language can be lexicographically ordered along an axis. Counting the words appearing in a document maps this axis into a long sparse vector that represents the document. Support vector machines allow to classify the points in this multidimensional space by drawing a hyperplane separating two categories. Document similarity can also be improved by mapping each word into the concept it represents (See W.2), so that the vector now contains concepts appearing in the document.

W.11. Latent semantic analysis. A well-known approach of corpora-based method is latent semantic analysis (Landauer, Dumais 1997), which applies singular value decomposition to reduce the dimensions of the term-document space, harvesting the latent relations existing between documents and between terms in large text corpora. Less computationally expensive techniques are based on mapping documents to a kernel space where documents that do not share any term can still be close to each other (Joachims et. al 2001; Liu et al. 2004). Another corpus-based technique that has been applied to estimate semantic similarity is PMI-IR (Turney, 2001). This information retrieval method is based on pointwise mutual information, which measures the strength of association between two elements (e.g., terms) by contrasting their observed frequency against their expected frequency.

W.12. As an evolution of Latent Semantic analysis, Latent Dirichlet Allocation (LDA) (Blei et al. 2003) proposes a generative approach, where each document is supposed to be "generated" as a *combination* of a given set of topics, in varying proportions. Further, each topic is considered as a probability distribution over words, and each word occurrence in a specific document is supposed to be attributable to a particular topic. Of course the problem is to find out, given a collection of documents, to find out the set of topics, for each topic its associated word probabilities, the topic responsible of generating each word, and the topic mixture for each document. This process is computationally costly, as it involves cycles through each and every word appearing in each document.

W.13. Frequently, topics of a collection of documents are computed by analyzing probabilistic distributions of terms (see, for instance, W.26) and in this way the topic model (a taxonomy, say) of the collection is obtained. (Ramírez and Brena, *this book*) observe that this approach is costly in computation time, and propose that a set of semantically similar documents be identified (as the answer to a *query*) as a semantic cluster, so that a collection of documents can be modeled as several of these semantic clusters, with overlaps. Their Query-based Topic Modeling framework (QTM) assumes the existence of a "preferred" set of queries, that cover most of the information in the set of documents and are somehow orthogonal (maximum semantic coherence). Two advantages of the QTM approach are 1) that it uses highly scalable clustering algorithms, and that 2) they can be executed using Map-Reduce parallelizing techniques.

W.14. (Zhihua Wei et al, *this book*) attack the problem found in a document: many words do not appear in such text (data sparseness). Therefore, it is useful to smooth the information, for instance to make use of statistic theory or linguistic information to assign a uniform distribution to absent words. Instead of doing this, the authors of this interesting chapter use Rough Sets to assign different values for absent words in different approximation regions.

W.15. (Makagonov and Reyes, *this book*) compare the similarity of two documents by reducing them to their *text document search image* (TDSI), which contains only words belonging to the domain of the text collection. The chapter defines: "A word belongs to the domain of a specific text collection if its relative frequency is higher – at least three times higher – than the frequency of the same word in the word frequency list of common language." In addition to finding similar documents, this chapter exploits the fact that the words in a document follow Zipf's law, in order to find restrictions and peculiar shapes on the plot (at the bi-logarithm scale) of the frequency of words for the collection under study. For instance, see its Figure 3.

W.16. (Garza and Brena, *this book*) use graph analysis to identify clusters of hyperlinks-connected documents, like the ones in the Wikipedia. The idea behind this method is to assume that in general a link could relate this document to another one with a similar topic (this is of course not the case of links to years in the Wikipedia). So they try to maximize the proportion of links that point to another document in the same cluster, and the resulting cluster is supposed to belong to a common topic. In order to avoid scalability problems, they make use of local analysis methods in the document graph.

Finding Similarity among Words or Symbolic Values

Instead of grouping documents, a basic approach is to group *words* into "semantically similar" if they represent or denote the same concept, or nearby concepts. This grouping can be carried out considering the words themselves (unattached to documents), considering the words contained in a group of documents (and the words in each document), or considering an a priori organization of words (a tree, say). Thus, we have two divisions:

- Grouping words using only word information (§1.2.1);
- Grouping words using a predefined taxonomy or arrangement of words (§1.2.2);
- Grouping words using a collection of documents (§1.2.3).

Intrinsic Relationship between Words

These methods group words using only word information.

W.17. Lemmatizers. Nouns, verbs and other types of words have declinations that express gender, tenses, number, etc. All these variations can be considered "similar" since they basically convey the same meaning (eat, ate, eaten), so it is useful to consider them as an equivalence class, and to select one of them (usually the verb in infinitive form, or the noun in singular) as their representative or *lemma.* Reducing derivatives to a single word is usually taken into consideration as preprocessing step, prior to the application of some of the methods we are presenting here (see for instance W.13).

W.18. Long sentences frequently appearing in a text are found ("maximal sequential patterns") using standard and new data mining algorithms by (García-Hernández *et al*, *this book*). Two algorithms are proposed, one for finding long sentences in a document, the other for finding them in a collection of documents. Both start with the Apriori algorithm (Agrawal & Srikant, 1994), and avoid frequent passes to the database by using a data structure residing in memory. They propose several uses of the maximal sequential patterns: document clustering, authorship attribution, question answering, mining hyponyms from the web, automatic text summarization.

W.19. Synonyms. WordNet is the classical example (Miller 1995), putting together words in groups of synonyms (Synsets), and then organizing these into a taxonomy.

W.20. An interesting variant is presented by (Hiram Calvo et al, *this book*). It analyzes the forms in which a verb and a subject can be related. They consider subject and object of a phrase as arguments to the verb, and answer the question of what arguments prefer which verbs. Is this a binary relation, one between three participants, or even more? One of the goals of this work is to identify situations where a verb argument is ill-chosen, like in "density *has brought me to you*", or in "*It looks like a* tattoo *subject*", by analyzing the plausibility of arguments.

Similarity of Words with Respect to a Given Taxonomy

These methods, known as "knowledge-based methods," group words using a previous arrangement or taxonomy of words.

W.21. Measuring the amount of information contained. (Resnik, 1995; Lin, 1998). In an information theoretic approach, the semantic similarity between two entities is related to their commonality and to their differences. Given a set of entities in a hierarchical taxonomy, the commonality of two entities can be estimated by the extent to which they share information, indicated by the most specific class in the hierarchy that subsumes both. Once this common classification is identified, the meaning shared by two entities can be measured by the amount of information needed to state the commonality of the two objects. Generalizations of the information theoretic notion of semantic similarity for the case of general ontologies (i.e., taxonomies that include both hierarchical and non-hierarchical components) have been proposed by Maguitman et al. (2005).

W.22. Distances in a concept network or taxonomy (Rada et al. 1989).

W.23. Taking into account both the distance (between words) in a taxonomy as well as the contained information (Jiang, Conrath 1997).

W.24. The semantic relatedness of two words could be assessed either by co-occurrence statistics over a large corpus, or by "knowledge-based methods" that rest upon known or given taxonomies or arrangements (WordNet, see 19, Wikipedia). Surely the features of two given words (as represented in the arrangement) contribute to ascertain their semantic similarity. But, how much each features contributes, and how? (Gentile, Zhang and Ciravegna, *this book*) find answers to these questions, using a random walk model on the (combined) features of the words. Weights are attached to these features.

W.25. Resting upon the concepts of abstraction and specialization (animal is more abstract than vertebrate; a specialization of vertebrate is mammal; mammal specializes into rodent, equines and other specializations), (Levachkine, Guzman 2007) organize words (they call them symbolic values) of a given topic in a hierarchy: a tree where each node is a word or, if it is a set, then the subsets of this set form a partition of it. The root denotes the most abstract concept; the leaves contain values that denote the most concrete items. On a hierarchy, it can be defined the function conf(r, s) that measures the confusion of using symbolic value *r* instead of the correct or intended symbolic value (word) *s.* In a chapter in this book, (Cuevas and Guzman), using *conf*, can compare the similarity of a collection of these values. The *inconsistency* of the set is found –it is large if the symbolic values are quite disparate, and is small if the values are quite similar. The *centroid* of the set is also found, it is their most likely representative, or *consensus value,* akin to the *lemma* of lemmatized words (See W.17). According to the context, a collection of values can have *several centroids,* and the chapter tells how to find them.

Grouping Words Guided by a Corpus of Documents

W.26. Using a proprietary version of the Latent Dirichlet Allocation method (Blei et al., 2003), (Magatti and Stella, *this book*) find topics in a document, and tag such documents according to a user-supplied taxonomy. Hierarchical clustering is used to select the optimal number of topics. The user needs to supply a set of similarity measures. The algorithm uses them for the topic labeling task. It also reviews several similarity measures between vectors (representing documents).

Finding Similarity among Concepts

An approach similar to §1.1.2 but more semantic in spirit consists in grouping the *concepts* (not the words) as "semantically similar" if they represent or denote the same objects in the real world, or near-by objects. Researchers have found out that a word can denote several concepts, according to the context or theme being described. (Wittgenstein used to say, humorously: "if words are ambiguous and concepts are unique, why don't we stop using words to communicate, and do it solely by concepts?"). Then, instead of grouping words similar to *java,* now we will group concepts similar to java1 (the programming language), Java2 (the island of Java) and java3 (a kind of coffee).

W.27. Formal Concept Analysis (Ganter, Wille 2004). Assume there are *n* properties that an object may or may not have (i. e., binary properties). A given object may have none, all, or more frequently, some of the *n* properties. These objects form a lattice, the top element represents the object(s) having *all n* properties, the bottom element is the object having *none* of these properties. Objects having one property (objects having "large"; objects having "red", …) lie near the bottom. A *concept* is any node in this lattice, for instance, node "red and heavy". Formally, a concept is a predicate of one variable, p(*x*), which is true if object *x* has property p (if it is red and heavy, say). The theory develops nicely, but the lattice is quite large. Even more, most of the nodes are useless, are chimeras (things that do not exist in the real world), for instance, object "flies and has scales and breathes air and has four legs."

W.28. (Kalfoglou, Schorlemmer 2002) uses FCA to establish a logical infomorphism between two local ontologies (ontologies of two different cultures –two islands, say; formally, a *local logic*). The mapping (the infomorphism) tries to preserve the relations (restrictions) found in each local ontology, and guarantees to preserve them for *normal instances*. It may be said that two concepts, one from ontology A, another from B, are equivalent if the infomorphism maps one into the other. It is a theoretical approach that perhaps has not yet found application.

W.29. (Olivares, Guzman 2004) develops a method (called COM) to compare concepts sitting in different ontologies (independently created) about the same or similar topic. His algorithm finds in an ontology the concept most similar to a given concept that belongs to another ontology. COM compares not only the names associated to both concepts, but their properties, their relations, their ascendants, etc.

CONCLUSION

We have reviewed in this introductory chapter some of the main proposals in this relatively new discipline of approximate, quantitative and "soft" methods for taking into account the meaning of pieces of text. So the main contribution of this book is to make a point about what is the current "state of affaires" in this novel area, and to present some of the methods in the voice of their authors.

It is our hope that in the next years, the advancement in the semantic analysis would be reflected in the creation of computer-based sophisticated systems for serving millions of users everyday goals, improving thus their lives in a meaningful way.

Ramon F. Brena
Tecnologico de Monterrey, Mexico

Adolfo Guzman-Arenas
Instituto Politécnico Nacional, Mexico

REFERENCES

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.), *Int. Conf. Very Large Databases* (pp.487-499). San Francisco: Morgan Kaufmann.

Antoniou, G., & Van Harmelen, F. (2004). *A semantic web primer*. Cambridge, MA: The MIT Press.

Blei, D. M., Andrew, N., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bollegala, D. (2007). Measuring Semantic Similarity Between Words using Web Search Engines. In *Proceedings of the Sixteenth International Conference on World Wide Web (WWW 07)*. (pp. 757-766). New York: ACM Press.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. [Cambridge, MA: MIT Press.]. *Computational Linguistics*, *32*(1), 13–47. doi:10.1162/coli.2006.32.1.13

Cuevas, A. D., & Guzman, A. (2008). A language and algorithm for automatic merging of ontologies . In Rittgen, P. (Ed.), *Handbook of Ontologies for Business Interaction* (pp. 381–404). Hershey, PA: IGI Global.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical central-ity as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*(1), 457–479.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11), 964–971. doi:10.1145/32206.32212

Ganter, B., & Wille, R. (2004). *Formal Concept Analysis*. Springer-Verlag.

Guzmán, A. (1998). Finding the main themes in a Spanish document. [Elsevier.]. *Journal Expert Systems with Applications*, *14*(1/2), 139–148. doi:10.1016/S0957-4174(97)00055-9

Guzman, A., & Olivares, J. (2004). Finding the Most Similar Concepts in two Different Ontologies. [Springer-Verlag.]. *Lecture Notes in Artificial Intelligence LNAI*, *2972*, 129–138.

Hatzivassiloglou, V., Klavans, J., & Eskin, E. (1999). Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP 99)*, pp. 203-212.

Haveliwala, T. H., Gionis, A., Klein, D., & Indyk, P. (2002). Evaluating strategies for similarity search on the web. In *Proceedings of the 11th international conference on World Wide Web*, pp.432-442. New York: ACM.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy, In *Proceedings of International Conference on Research in Computational Linguistics (ROCLING X)*, Taiwan, pp. 19–33.

Joachims, T. (2002). Learning to classify text using support vector machines: Methods, theory, and algorithms. *Computational Linguistics*, *29*(4), 656–664.

Joachims, T., Cristianini, N., & Shawe-Taylor, J. (2001). *Composite kernels for hypertext categorization* (pp. 250–257). MACHINE LEARNING-INTERNATIONAL WORKSHOP.

Kalfoglou, Y., & Schorlemmer, M. (2002). Information-Flow-based Ontology Mapping. In *Proceedings of the 1st International Conference on Ontologies, Databases and Application of Semantics (ODBASE '02)*, Irvine, CA, USA.

Ko, Y., Park, J., & Seo, J. (2004). Improving text categorization using the importance of sentences. [New York: Elsevier.]. *Information Processing & Management*, *40*(1), 65–79. doi:10.1016/S0306-4573(02)00056-0

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*(2), 211–240. doi:10.1037/0033-295X.104.2.211

Lapata, M., & Barzilay, R. (2005). *Automatic evaluation of text coherence: Models and representations.* IJCAI 2005 Proceedings, pp. 1085.

Levachkine, S., & Guzman-Arenas, A. (2007). Hierarchy as a new data type for qualitative variables. [Elsevier.]. *Expert Systems with Applications*, *32*(3), 899–910. doi:10.1016/j.eswa.2006.01.024

Lin, D. (1998). An Information-theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 98)*, pp. 296-304, San Francisco, CA.

Liu, Y., & Zong, C. (2004). Example-based chinese-english MT, Systems, Man and Cybernetics, 2004 *IEEE International Conference on, 7*, pp.6093-6096. Washington, DC: IEEE Press.

Lorenzetti, C. M., & Maguitman, A. G. (2009). A semi-supervised incremental algorithm to automatically formulate topical queries. [Elsevier.]. *Information Sciences*, *179*(12), 1881–1892. doi:10.1016/j.ins.2009.01.029

Maguitman, A., Menczer, F., Roinestad, H., & Vespignani, A. (2005). Algorithmic Detection of Semantic Similarity. In *Proceedings 14th International World Wide Web Conference (WWW 05)*, pp. 107 - 116, New York: ACM Press.

Manning, C., & Schutze, H. (1999). Word Sense Disambiguation . In *Foundations of Statistical Natural Language Processing* (pp. 229–261). Cambridge, MA: MIT Press.

Marshall, C. C., & Shipman, F. M. (2003). Which semantic web? In *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pp. 57-66. New York: ACM Press.

Miller, G. A. (1995). WordNet: a lexical database for English. [New York: ACM Press.]. *Communications of the ACM*, *38*(11), 39–41. doi:10.1145/219717.219748

Olivares-Ceja, J. M., & Guzman-Arenas, A. (2004). Concept similarity measures the understanding between two agents. In *Natural Language Processing and Information Systems*. Lecture Notes in Computer Science, LNCS 3136, pp. 366-376, Springer.

Pierrakos, D., Paliouras, G., Papatheodorou, C., Karkaletsis, V., & Dikaiakos, M. (2003). *Construction of Web community directories using document clustering and Web usage mining*. In ECML/PKDD 2003: First European Web Mining Forum proceedings.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. [Washington, DC: IEEE.]. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 17–30. doi:10.1109/21.24528

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI 95)*, pp. 448-453.

Sahami, M., & Heilman, T. (2006). A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets. In: *Proceedings of the Fifteenth Inter-national World Wide Web Conference (WWW 06)*, pp.377-386. New York:ACM Press.

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML 01)*, pp.491-502. San Francisco: Morgan Kaufmann.

Yates, R. B., & Neto, B. R. (1999). *Modern information retrieval*. New York: Addition Wesley.

KEY TERMS AND DEFINITIONS

Clustering: Action of grouping together related items in a collection; it is considered as a non-supervised form of learning.

Corpus-Based: Any method relying entirely on the actual contents of a collection of documents.

Infomorphism: Mapping among information structures.

Knowledge-Based: Any method supported by knowledge previously acquired by humans.

Quantitative: Related to measures, as opposed to qualitative.

Semantics: Related to the meaning of languages, phrases, etc.

Taxonomy: Tree-like classification.