Automatic Annotation of Cyttron Entries using the NCIthesaurus

David Graus

Media Technology MSc,
Leiden Institute of Advanced Computer Science
Leiden Universiteit, Leiden
david@graus.nu

Abstract

Semantic annotation uses human knowledge formalized in ontologies to enrich texts, by providing structured and machine-understandable information of its content. This paper proposes an approach for automatically annotating texts of the Cyttron Scientific Image Database, using the NCI Thesaurus ontology. Several frequency-based keyword extraction algorithms were implemented and evaluated, aiming to extract important concepts and exclude less relevant ones. Furthermore, topic classification algorithms were applied to identify important concepts which do not occur in the text. The algorithms were evaluated by comparison to annotations provided by experts. Semantic networks were generated from these annotations and an ontology-based similarity metric was applied to perform the comparison. Finally the networks were visualized to provide further insights into the differences of the semantic structure generated by humans, and the algorithms.

Tags: Semantic annotation, ontology-based semantic similarity, keyword extraction, topic classification, network visualization

1. Introduction

Semantic annotation is the process of enriching a document by attaching concepts to it, providing extra metadata and context in a structured and machine-understandable way. For automated semantic annotation, an ontology is often used. This is a semantic repository with a set of concepts and their relationships for a specific domain. The Cyttron Scientific Image Database for Exchange (CSIDx) consists of images manually described and annotated by scientists. This paper explores an approach of automatically annotating the textual descriptions of this image database with concepts from the NCIthesaurus ontology.

The proposed approach tries to achieve this in three steps. The first is keyword extraction. The goal of this step is to create a representation of a Cyttron entry description that includes the most relevant concepts, while excluding irrelevant words. This is attempted by implementing and evaluating a variation of frequency-based keyword extraction algorithms. The second step is semantic annotation. It is approached in two ways: by extracting literal occurrences of ontology concepts from the keyword representations of the Cyttron entry, and by a topic classification approach which aims to retrieve relevant concepts which do not occur in the Cyttron entry. The third and final step consists of evaluating the various annotations by comparing them to expert provided annotations. This comparison is done in two ways: by confusion matrices, to measure if and how the automatically generated annotations overlap with the expert annotations, and by a path-based semantic similarity metric, which uses the semantic information contained in the hierarchical structure of the NCIthesaurus ontology to provide a more detailed comparison. In this final evaluation step the generated semantic networks are visualized in order to provide insights that might have otherwise been overlooked.

This paper starts by describing related work in the areas of frequency-based keyword extraction, topic classification, ontology-based semantic similarity measures and network visualization in section 2. In section 3 the methods used in this study will be explained. In section 4 the results and observations will be presented, and they will be discussed in section 5. Finally the conclusion will be presented in section 6.

2. Theory

There are various approaches to keyword extraction. In this study a frequency-based approach was chosen, as there is no suitable standard corpus in the biomedical domain for the particular area of free-form text descriptions by scientists. As opposed to corpus linguistic approaches to keyword extraction, frequency-based keyword extraction does not require an extensive corpus. Frequency-based keyword extraction assumes that frequently occurring terms in a document are important [1]. This approach is commonly expanded by the assumption that terms are more important when they appear frequently in a single document, but infrequently in the document's context: the entire corpus of documents [11]. This weighting is called Term Frequency – Inverse Document Frequency (TF-IDF) weighting. The efficiency of TF-IDF weighting in keyword extraction is highly dependent on the quality and size of the corpus. A large corpus is needed to provide a reliable IDF-metric [14]. Therefore it is not suitable for this particular application, where the corpus is too small to provide a reliable IDF metric.

Automated topic classification has seen a surge in popularity in the field of Information Retrieval, with the increased availability of digital documents [19]. A common approach for topic classification is to convert an input document and an index of possible 'topics' to a vector, using a vector space model. The vector of the input document can then be compared to each index document, to retrieve similar documents. The most simple approach to convert a document to vector space is by using the bag of words model. In the bag of words model, a document is treated as an unordered list of words. The document's vector consists of a list of frequencies of those words. As with frequency-based keyword extraction, the TF-IDF function is commonly applied to assign importance to words. Furthermore, it is common to remove common stop words, and stem remaining words before converting the document to its vector-space representation [19].

A distinction of two approaches in ontology-based similarity measurement can be made: information content (IC) based measures and path-based measures [5]. IC measures depend on a corpus to provide information on relative word frequencies and word co-occurrences. A path-based measure was chosen, as there is no standardized biomedical corpus available and the Cyttron database is considered too small. There are several commonly applied path-based similarity measures [12, 20, 15]. These path-based measures are frequently designed for WordNet, but used and validated in a number of other domains, particularly the (bio)medical domain [2, 3, 4, 16, 17]. In this paper Leacock & Chodorow's semantic similarity metric is implemented, as in the biomedical domain this measure in several cases outperforms, even if by a small margin, similar path-based approaches [2, 4].

The discipline of network visualization in a scientific context dates back as far as 1736 [13]. Two different perspectives can be distinguished in network visualization: the mathematical perspective of graph theory and graph drawing, which deals primarily with layout algorithms for complex networks. And the perspective of information visualization, where visual design principles are applied to optimize the communication of information. There is no universal language in network visualization [13], but basic design principles e.g. Gestalt theory, can be used in the context of network visualization. Ontology visualization has been extensively studied, it deals primarily with how to present large quantities of information [8], and thus combines graph drawing and information visualization aspects. But as this study does not include the visualization of ontologies as a whole, but rather their concepts without the ontologies' hierarchical context and structure, this particular subfield is less relevant.

3. Method

3.1 Keyword Extraction

Different variations of standard frequency-based keyword extraction algorithms were implemented: most frequent words, most frequent nouns (after part-of-speech tagging), and most frequent bi- and trigrams (small phrases consisting out of two or three words which frequently co-occur), and one which combines all four keyword extraction algorithms. Next to these keyword extraction algorithms, the original, unedited Cyttron entry is used. This brings the total to six different representations:

1. Literal: Use the raw text as-is.

2. Most frequent words: Remove common stop-words, return n most frequently occurring words
3. Most frequent nouns: Remove common stop-words, identify each word's part of speech (noun, verb, etc.) using NLTK's POS-tagger, extract identified nouns, return n most frequently occurring nouns
4. Most frequent bigrams: Return n most frequently occurring bigrams
5. Most frequent trigrams: Return n most frequently occurring trigrams
6. Keyword combo: Sum of the results of method 2-5

These six different representations are reapplied in three more variations: one which includes Porter stemming of each extracted keyword, and one which includes synonym generation for each keyword using WordNet, and finally one which includes both stemming and synonym generation. This brings the total amount of keyword extraction algorithms to 24.

3.2 Semantic Annotation

For the actual annotation, the NCIthesaurus is hosted in a local triple store, accessible through a SPARQL endpoint. A list of concept label-URI pairs is created to use during the annotation task. Two different annotation methods are applied.

3.2.1 Annotation I – Literal Occurrences

The first semantic annotation method uses literal occurrences of ontology concepts. The NCI Thesaurus includes 89.129 concepts. For each Cyttron entry the occurrences of concept labels are counted. Word boundaries are taken into account to avoid the word 'Epithalamus' matching the concept 'Thalamus'. It was attempted to include synonyms, bringing the total amount of labels to 258.051. But as these are so numerous and frequently overlap (particularly in abbreviations), the amount of concepts increased by a large margin. This would mean the data needed the extra challenge of disambiguation.

3.2.2 Annotation II – Topic Classification

The second approach of semantic annotation uses topic classification to identify relevant topics which do not occur in the Cyttron entry. For this goal, an index is created, consisting of the 50.804 NCIthesaurus concepts which are supplied with a definition. Both input document (the Cyttron entry) and index documents are then converted to vector space. This conversion is achieved in two steps: first, the document is converted to vector space using the Bag of Words model: a list of frequencies of important features (words). Next, this Bag of Words representation is transformed by TF-IDF weighting, which assigns weights to the features based on their occurrence in the index document and the whole index. The resulting TF-IDF weighted vectors are then compared using the Cosine similarity measure.

To identify the important features used for the bag of words conversion, the BioMedCentral OpenAccess full-text corpus[1] is used. This corpus consists of over 100.000 published BioMedCentral research articles. It was chosen because of its domain and large size. This corpus is considered suitable in the task of indexing NCIthesaurus definitions, as opposed to only the free-form textual descriptions of the Cyttron entries, as would be the case if a TF-IDF approach was chosen for the frequency-based keyword extraction.

To identify features in the BioMedCentral corpus, it is first preprocessed by:

1. Extracting the main article body of each article
2. Filtering 'blank' entries (*"For the full text, download the PDF"*)
3. Removing punctuation, common stop words and non-alphanumeric content
4. Applying the Porter-stemming algorithm to stem the remaining words

This leaves a corpus with 99.432 documents, containing 1.136.471 unique features. These features are then added to a dictionary to use for the Bag-of-Words transformation. In this case, each document is represented by a 1.136.471-dimensional vector.

Five different topic classification algorithms are distinguished by varying the cut-off point (or 'tolerance') of the returned concepts:

1. Return any concept over 75% similar
2. Return any concept over 90% similar
3. Return best 5 concepts
4. Return best 10 concepts
5. Return 10% best concepts

[1] http://www.biomedcentral.com/about/datamining

3.3 Evaluation

To validate the annotations produced by the annotation methods, expert data was acquired. An online survey was set up and three biomedical experts were invited to manually annotate a selection of eight Cyttron entries. These entries were manually selected, being characterized by having high quality images, descriptive annotations, varying lengths of descriptions and representative topics[2]. The survey was set up so that the experts were provided with the image, the Cyttron entry description and the ontology concepts with their definitions (where applicable), presented through the BioPortal ontology viewer[3].

To further validate the chosen approaches a set of random NCIthesaurus concepts is generated. These annotations were created by randomly picking between 2 and 10 concepts for each Cyttron entry[4].

Finally, several sets of annotations were collected. Each set of annotations consists out of eight separate subsets, one for each selected Cyttron entry. This resulted in the following annotation sets:

- Three sets by the experts[5]
- 24 sets from Annotation method I
- Six sets from Annotation method II
- One set of randomly generated annotations

3.3.1 Confusion Matrices

The first step of evaluation is by measuring the performance of each algorithm-produced annotation set by comparing it to the expert annotation sets. A concept which appears in an expert annotation is classified as positive, the remaining concepts are classified as negative. For each algorithm the following variables were measured:

1. How many concepts were included by both the algorithm and the expert (correctly classified as positive)
2. How many concepts were included by the algorithm, but excluded by the expert (incorrectly classified as positive)
3. How many concepts were excluded by both the algorithm and the expert (correctly classified as negative)
4. How many concepts were excluded by the algorithm, but included by the expert (incorrectly classified as negative)

With this data subsequently the proportion of correct predictions (Precision), the fraction of correct annotations that were retrieved (Recall), and the proportion of total correct annotations (Accuracy) is calculated for each algorithm.

3.3.2 Semantic Similarity

Path-based semantic similarity measures can measure the similarity between two concepts in an ontology, using its hierarchical structure. Path-based similarity is based on the assumption that concepts that are closer in an ontology's hierarchy are semantically more similar. In this study, Leacock & Chodorow's semantic similariy measure is used. This measure scales the shortest path between two concepts by the maximum depth of the ontology, and includes log-smoothing [16]:

$$sim(a,b) = -\log(Np/2D)$$

Where the similarity between node a and node b is given by dividing the number of nodes in shortest path p (Np) through the maximum depth D*2 [9]. The maximum depth of the NCIthesaurus is 15[6].

Shortest path between nodes

To retrieve the shortest path between two concepts, a SPARQL-driven Breadth-First-Search algorithm is implemented. The algorithm takes a start and target node, retrieves the shortest path by passing through each node and querying its neighbors until it finds the target. As this method can take several hours to find a path between two distant concepts due to the size of the ontology, the path is stored in a local database for later reference. It is important to note that path-length is measured only by taking into account structural relations, more specifically, only the 'is_a' and 'part_of' relation [17]. In the OWL representation of the NCIthesaurus, used in this study, the is_a relation is modeled by the "rdfs:subClassOf" predicate, and five distinct part_of relations are defined.

The NCIthesaurus is subdivided in 20 clusters. Each cluster represent a different conceptual entity e.g. "Disease, Disorder or Finding", "Gene" and "Biological Process". As the concepts in these clusters share no 'cross-cluster' is_a or part_of

[2] Table 3.1
[3] http://bioportal.bioontology.org/ontologies/1032/
[4] Table 3.3
[5] Table 3.2
[6] http://bioportal.bioontology.org/ontologies/1032/

relations, the algorithms first determine whether both nodes are located in the same cluster. If they are not, the shortest paths from each node to the root concept is calculated, and summed to provide the shortest possible path between the nodes.

Subgraph Similarity

To measure the similarity between two sets of annotations, the sets are represented as subgraphs by treating the concets as nodes in the NCIthesaurus graph. For each node n in set 1, the shortest paths to each node in set 2 is measured. This shortest path is stored for each node n. When all nodes in set 1 have been processed, the process is repeated from nodes in set 2 to set 1, omitting duplicate paths. This approach was chosen to make sure each node is included in the subgraph similarity measurement.

3.4 Network Visualization

The networks visualized in this study aim to increase understanding of the semantic similarity of two separate subgraphs in a single ontology. After performing the similarity measurement, hierarchical information is discarded, and only the concepts and their similarity is displayed.

For effective information visualization it is important to carefully choose which information to communicate and how. In the case of this study, two properties are considered most important: the similarity between two nodes, and the node's source: whether the node comes from an automatic annotation or an expert annotation. Furthermore, other properties contribute to a better understanding of the differences and similarities between the subgraphs of experts and algorithms, e.g. node's type (which cluster it originates from) and node's specificity (concept 'depth'). These two properties were selected because they contain important semantic information, which also indirectly influence the similarity measurement: less specific concepts are more likely to have short paths, as they are close to the root and make long cross-cluster paths unlikely, and different types of concepts increase the likeliness of lower semantic similarity, as different clusters are connected through the root node.

In order to map these properties to the representation of the network, various visual properties of nodes and edges can be altered, e.g. color, shape, size, orientation, position [13]. As the node's similarity is the core business, it should have a strong emphasis in the visualization. Grouping is very powerful property in network visualization [13]. This mechanism is described by Gestalt theory's law of proximity. Intuitively, it makes sense to group similar nodes. By including besides the semantic similarity between nodes from different subgraphs, also each node's similarity *within* the subgraph, similar concepts can be automatically drawn into proximity by creating a force-directed graph, where the similarity between two nodes is mapped to the edge's attraction force. Similar concepts are thus pulled together stronger than less similar concepts. Additionally, the brightness of the edges is adjusted according to its strength: making strong edges brighter than weaker edges, further emphasizing the visual distinction of high and low similarity.

Another core element in the comparison of annotation sets is the source of a concept. It is mapped to the node's color to allow similar grouping, not by proximity but by shared color.

As the depth of an NCIthesaurus' concept is a relevant factor in the semantic similarity calculation, it was mapped to the size of the node, drawing more specific nodes larger than less specific ones.

Furthermore, the concept type in the annotation can provide extra information of the content of a document. Each type was assigned a distinct shape, allowing grouping by shape.

4. Results

4.1 Confusion Matrices

4.1.1 Annotation method I

The top three performing algorithms are consistent for all three expert annotations: freqWords, nounWords and stemmed nounWords [7]. It is interesting to note the performance of the nounWords algorithm, as a POS-tagger which was untrained for this specific domain was used. WordNet variants always score relatively low, with the raw entries in their stemmed and regular way consistently worst and significantly worse than other algorithms.

[7] Figure 1.1.1-1.1.3

It is important to note that due to the large number of negatives (all concepts which are not included in the expert's annotation) the Accuracy number is skewed, with each algorithm scoring accuracy of over 99.8%. The high accuracy is not meaningful in its absolute sense, but it is meaningful in comparison to the other accuracies.

4.1.2 Annotation method II and Random Annotations

None of the Annotation II algorithms, nor the random annotations got a single annotation right with any of the expert annotations.

4.1.3 Expert Annotations

Expert 1 and Expert 3 have 9 overlapping concepts in their annotations, out of a total of 67 annotated concepts. There is no overlap between Expert 2 and the either Expert[8].

4.2 Ontology-based Annotation Analysis

4.2.1 Specificity of Concepts

The specificity of an annotated concept is determined by the distance of its shortest path to the root. The average specificity of the expert annotations[9] differ substantially, with an average specificity of 4.09, 1.92 and 5.93 for Expert 1, 2 and 3 respectively. Annotation method I[10] has an average depth roughly between 3 and 4. The algorithms of Annotation method II tend to be more specific[11].

4.2.5 Types of Concepts

By retrieving the closest node to the root for each annotated concept, the distribution of types of concepts per for each annotation can be analyzed.

Annotations produced by Annotation method I consistently have a very high amount of "Property or Attribute" concepts[12]. The experts tend to annotate "Anatomic Structure, System or Substance" concepts most frequently[13]. Another observation with this data is that most Annotation algorithms include a large number of concepts, frequently surpassing 100 annotations per Cyttron entry by a large margin. Experts annotate on average between 4-5 concepts per entry.

The results further show that the stemming the Cyttron entries results in a higher number of concepts, which is to be expected as plural words are only retrieved in Annotation method I after being stemmed.

4.2.2 Semantic Similarity between annotations

Using Leacock & Chodorow's path-based semantic similarity metric, two sets of annotations are compared by measuring the shortest distance between the subgraphs created from both annotations. For reference, the semantic similarity between experts was measured[**14**]. These averages lie between 1.75 and 1.98. It is interesting to note that the standard deviation between Expert 1 and 3 is relatively high. This could be explained by the fact that the annotations by Expert 2 are relatively close and unspecific topics.

When comparing the results of Annotation method I to Expert 1[**15**], the best performing approaches are all stemmed, either bigrams, combo or literal representations, with average similarities of 1.87, 1.86 and 1.85 respectively. These similarities are close to those of the inter-expert similarities. The performance of the literal algorithms stand out: where they perform bad in the confusion matrix evaluation, they seem to produce more semantically similar annotations than keyword representations of the entries.

Interestingly, the best performing Annotation I algorithms generate more semantically similar annotations to Expert 2[**16**], than expert 1 and 3 do. Annotation I algorithms produce annotations with low similarity to those of Expert 3: highest similarities are between 1.75 and 1.79[**17**].

The best performing algorithms for each expert vary heavily, and the similarity scores of these best annotations are generally lower than those of the best approaches in Annotation I[**18**].

Annotation method II had no direct matches with any of the expert annotations in the confusion matrix evaluation. The semantic similarity between

[8] Table 1.2.1
[9] Table 2.1.1
[10] Table 2.1.2
[11] Table 2.1.3
[12] Table 2.2.2
[13] Table 2.2.1
[14] Table 2.2.1
[15] Table 2.2.2.1
[16] Table 2.2.2.2
[17] Table 2.2.2.3
[18] Table 2.2.3.1-2.2.3.3

the annotations produced by this method and the experts is lower than between the annotations created by Annotation method I and the experts. The random annotations score lower than the best performing annotations of method II, for expert 1 and 3[19]. Surprisingly, the random annotations have a higher similarity to the annotations by expert 2, than with the annotations generated by method II.

4.2.3 Semantic Similarity per Expert Annotation

In an attempt to further specify features of expert annotations, the semantic similarity within an annotation is calculated[20]. It is important to note that these numbers cannot be directly compared to the semantic similarity numbers between annotations: in this case, all concepts in the annotation were cross-compared, instead of picking the shortest paths from a node in set 1 to a node in set 2. Again, Expert 2 is significantly different from the other experts; there is more similarity between the concepts used in each annotation set. This can be explained by the relative low specificity of the annotations.

4.3 Network Visualization

The visualizations allows us to confirm some of the observations described in the previous section. Figure 1 shows the high number of 'Property or Attribute' annotations of the annotation produced by method I (red circles), and how they are unevenly distributed: in some cases they have a relatively high similarity to concepts annotated by the expert (blue).

Furthermore, in some cases clusters around certain topics can be distinguished in the expert annotations, for example the cluster of brain-related concepts in Figure 2. In Figure 3, the low specificity of Expert 2's annotation is clear by the small size of shapes.

5. Discussion

The accuracy of literal representations of Cyttron entries is significantly lower in the confusion matrix evaluation. The literal representation include more annotations per Cyttron entry, increasing the False Positive rate and thus decreasing accuracy. In this regard, the keyword extraction algorithms are effective. The WordNet representations scored low for likely the same reason: the algorithms produce a much higher amount of annotations, since for each word multiple synonyms can be generated. Surprisingly, the top performing algorithms are quite consistent: most frequent words, nouns, bigrams and trigrams, stemmed or not. These algorithms are characterized by containing a relatively low number of concepts per annotation. As the nounWord extraction seems to perform rather well, it might prove worthwhile to train a POS-tagger for this specific domain.

An important observation is that there does not seem to be a general 'agreement' between the three expert annotations in both evaluations. In the confusion matrices, there were merely 9 concepts overlapping between Expert 1 and 3. The average specificity of the annotations differ greatly, and there is no high semantic similarity between the annotations. In fact the similarity between experts is quite comparable to the similarity between best performing algorithms and expert annotations, in both annotation methods.

As opposed to the evaluation by confusion matrix, in some cases literal annotations outperform keyword representations in the semantic similarity measure. This is caused by the fact that similarity between two annotation sets is not negatively impacted by the total amount of annotations.

When looking at the type of concepts in the annotation, it becomes clear the algorithms include a high proportion of 'Property or Attribute' type concepts. This proportion is smaller in the expert annotations, who were more likely to annotate concepts of the type "Anatomic Structure, System, or Substance". Intuitively this makes sense, considering the texts that were annotated were descriptions of microscopic images. Increasing the list of stopwords to include some of these 'Property or Attribute' words could increase the accuracy of the algorithms. In this study, a quite limited list of stop words is used.

Annotation method I has proven to be the best approach for this domain: both in confusion matrices and semantic similarity, this method scores higher than Annotation II and the set of random annotations. While annotation method II did not produce any 'direct hits' with the expert annotations, it did produce more semantically similar annotations than the random annotations. Except for the annotations of Expert 2. In this case,

[19] Table 2.2.4
[20] Table 2.2.5

the higher similarity of the random annotations can be partly explained by considering the average specificity of the annotations of Expert 2: lower specificity means the concepts are close to the root node, which makes long crossing-cluster paths less likely.

When it comes to the observations concerning semantic similarity a side note has to be made: as these similarities are purely based on the length of paths between concepts, they do not allow us to make assumptions about the quality of the annotations. For example, concepts of the type "Property or Attribute", can still provide a high similarity, even if they are considered irrelevant by experts, by being excluded in the annotations.

The second side note is that the high amount of annotated concepts per entry, decreases performance in the confusion matrix evaluation, but is ignored by the semantic similarity measurement, since only the closest paths between concepts from two sets are considered.

Looking at the average specificities of the algorithm's annotations, particularly annotation method I, could lead to conclude that similarities are in fact not very high, as the annotations are generally less specific than the expert annotations.

In this study, these attributes have been analyzed in isolation, it could prove to be a worthwhile approach to combine these, to produce a better set of heuristics for automatic semantic annotation.

The visualizations provide more insights into the semantic similarity, as the different types of concepts are easily recognized. Furthermore, it can provide insights to important topics in the Cyttron entries, by identifying clusters of similar annotated concepts.

6. Conclusion

Annotation method I seems to be the best performing annotation approach in this study. Keyword extraction has proven to be a sensible step, as the keyword representations of the Cyttron entries commonly outperform the literal entries.

Even if Annotation method II did not produce any directly matching annotations, it has shown to produce more semantically similar annotations than a set of random annotations.

By analyzing the hierarchical and structural aspects of the annotations in the ontology, some characteristics of expert annotations and annotations generated by the proposed methods have been identified: the strong emphasis on 'Property or Attribute'-type concepts with Annotation method I, the varying specificity and similarity between expert annotations. Of course, more extensive expert data might have provided the opportunity to discover common characteristics of human annotations.

The path-based similarity measurement has shown that it is able to provide more detailed information on algorithm performance. It shows that Annotation method II produces more similar annotations than the random annotations, a distinction lost in the confusion matrix evaluation, where both algorithms produce identical results. However, since this measure does not take the amounts of annotated concepts into account, annotations with a large number of concepts are not sufficiently 'punished'.

An advantage of the approach chosen in this study lies in its domain independence. As only the topic classification step uses an external corpus, the method can be used in any domain where a sufficiently descriptive ontology is available. Furthermore, it is possible to use the index itself to identify important features, instead of an external corpus.

The approach of using sub-graph similarity to determine semantic similarity between annotations, in this study used to provide more information of the annotations, could be further studied to apply in different tasks, e.g. document similarity or summarization. If a consistent approach of keyword extraction and automatic annotation is chosen, path-based measures could prove useful in determining the likeliness or core content of multiple documents.

7. References

1. AggarWal, C. & Zhai, C. "MINING TEXT DATA" Kluwer Academic Publishers.
2. Al-Mubaid, H., & Nguyen, H. A. (2006). A cluster-based approach for semantic similarity in the biomedical domain. Conference Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society, 1(Ic), 2713-2717. IEEE Computer Society.
3. Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics, 44(1), 118-125. Elsevier Inc.
4. Batet, M., Sánchez, D., Valls, A., & Gibert, K. (2010). Exploiting Taxonomical Knowledge to Compute Semantic Similarity: An Evaluation in the Biomedical Domain. Knowledge Creation Diffusion Utilization, 6096/2010, 274-283.
5. Blanchard, E., Harzallah, M., Briand, H., & Kuntz, P. (2005). A typology of ontology-based semantic measures. EMOIINTEROP'05 Proc Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability (Vol. 5). Citeseer.
6. Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Evaluation,2(12), 29-34. Citeseer.
7. Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics, 32(1), 13-47. MIT Press.
8. Dmitrieva, J. (2011). Aspects of Ontology Visualization and Integration. PhD Thesis.
9. Fellbaum, C. (1998). Wordnet – An Electronic Lexical Database. MIT Press.
10. Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-Specific Keyphrase Extraction. Proceedings of the 16th International Joint Conference on Aritificial Intelligence IJCAI (Vol. 16, pp. 668-673). Citeseer.
11. Hulth, A., Karlgren, J., Jonsson, A., Bostrom, H., & Asker, L. (2001). Automatic Keyword Extraction Using Domain Knowledge. (A. Gelbukh, Ed.)English, 2004, 472-482. Springer.
12. Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), WordNet An Electronic Lexical Database (Vol. 49, pp. 265-283). MIT Press.
13. Lima, M. (2011). Visual Complexity – Mapping Patterns of Information. Princeton Architectural Press.
14. Liu, Feifan, Pennell, D., Liu, Fei, & Liu, Y. (2009). Unsupervised approaches for automatic keyword extraction using meeting transcripts. Proceedings of Human Language Technologies The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on NAACL 09, (June), 620. Association for Computational Linguistics.
15. Pedersen, T. Patwardhan, S. Michelizzi, J. 2004. WordNet::Similarity: measuring the relatedness of concepts. In Demonstration Papers at HLT-NAACL 2004 (HLT-NAACL--Demonstrations '04). Association for Computational Linguistics, Stroudsburg, PA, USA, 38-41.
16. Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics, 40(3), 288-299.
17. Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. (P. E. Bourne, Ed.)PLoS Computational Biology, 5(7), 12. Public Library of Science.
18. Rehurek, R., & Sojka, P. (2004). Software Framework for Topic Modelling with Large Corpora. Processing (pp. 45-50). ELRA.
19. Sebastiani, F. (2001). Machine Learning in Automated Text Categorization.ACM Computing Surveys, 34(1), 1-47. ACM.
20. Wu, Z., & Palmer, M. (1994). Verb Semantics and Lexical Selection. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, 6. Association for Computational Linguistics.