Conceptual Hierarchical Clustering of Documents using Wikipedia knowledge

Gerasimos Spanakis, Georgios Siolas and Andreas Stafylopatis

Intelligent Systems Laboratory
School of Electrical and Computer Engineering
National Technical University of Athens
15780, Zografou, Athens, Greece

Abstract. In this paper, we propose a novel method for conceptual hierarchical clustering of documents using knowledge extracted from Wikipedia. A robust and compact document representation is built in real-time using the Wikipedia API. The clustering process is hierarchical and creates cluster labels which are descriptive and important for the examined corpus. Experiments show that the proposed technique greatly improves over the baseline approach.

1 Introduction

Nowadays, Wikipedia has become one of the largest knowledge repositories with many advantages (size, dense link structure between articles, brief anchor texts e.t.c). This paper introduces an efficient Conceptual Hierarchical Clustering (CHC) technique of documents, using a document representation based on Wikipedia knowledge and exploiting Wikipedia article features (ingoing/outgoing links etc.) Clusters produced have labels, informative of the content of the documents assigned to each specific cluster.

2 Related work

There has been a growing amount of research in ways of enhancing text categorization and clustering by introducing Wikipedia external knowledge [3], [1]. Gabrilovich and Markovitch [3], propose a method to improve text classification performance by enriching document representation with Wikipedia concepts. Banerjee et al. [1] extend the method applied in [3] by using query strings created from document texts to retrieve relevant Wikipedia articles. Both methods only augment document representation with Wikipedia concepts content without considering the hierarchical structure of Wikipedia or any other features of the ontology. All of the papers mentioned above, rely on existing clustering techniques (mostly k-Nearest Neighbors and Hierarchical Agglomerative Clustering) whereas in this paper we extend the idea of [5] and introduce a novel clustering technique, Conceptual Hierarchical Clustering (CHC).

3 Document Representation Model using Wikipedia

Our goal is to extract Wikipedia concepts which are described by one or more consecutive words of the document. In our approach, we overcome the bottleneck of extracting all possible N-grams, by choosing to annotate each document's text with Part-of-Speech information using the TreeTagger tool provided by [8]. Wikipedia articles have descriptive titles, so it is not necessary to perform stemming or remove stop words during document preprocessing. After this procedure, we keep those consecutive words which are nouns and proper nouns (singular or mass or plural) along with prepositions, subordinating or coordinating conjunctions and the word *to* (POS tags in the Penn Treebank Tagset [6]). By grouping consecutive words with the previous POS tags we perform full *Noun Phrase* extraction, forming our candidate concepts.

For each candidate concept, we automatically check "on-the-fly" whether it exists or not as a Wikipedia article using the Wikipedia API. If the concept has multiple senses (so there are multiple Wikipedia articles referring to the same Noun Phrase), we use the disambiguation technique proposed by [2] in order to choose the most appropriate sense. Once we obtain a unique mapping between the candidate concept and Wikipedia, the concept is selected as a component of the document vector which is about to be formed. At the same time, using the Wikipedia API, for every selected concept i, we extract the features presented below :

- $Content_i$: the corresponding Wikipedia article text
- $Links_i$: links from the corresponding article to other articles
- $BackLinks_i$: articles which have a link to the examined article
- $PageHits_i$: the articles in which the examined article (Noun Phrase) is simply present, either as link or not (plain text)

After the extraction of the features mentioned above for every concept i in a document j, we combine them with the original document features, as described in the equations below, in order to form a richer document representation.

- Weighted Frequency ($Wfreq$) is defined by :

$$WFreq_{j,i} = size_i * frequency_{j,i} \tag{1}$$

  where : $size_i$ is the number of words that form concept i and $frequency_{j,i}$ stands for how many times concept i occurs in document j.

- *LinkRank* is a measure of how many links a concept has in common with the total of those contained in a document, thus it is a measure of the importance of the concept to the document and is formally defined as :

$$LinkRank_{j,i} = \frac{|Links_i \bigcap Links_{Doc_j}|}{|Links_{Doc_j}|} \tag{2}$$

  where : $Links_i$ is the set of Links of concept i and $Links_{Doc_j}$ is the set of Links of document j, defined as all the links of all concepts that represent document j.

- $ConceptSim$ is the similarity between the document and the article text of a concept contained in the document, computed in the classic term frequency - inverse document frequency ($tf - idf$) vector space, which is given by the following equation :

$$ConceptSim_{j,i} = \cos(\mathbf{v}_j, \mathbf{v}_i) \tag{3}$$

  where : $\mathbf{v}_j$ is the $tf - idf$ vector of document j, $\mathbf{v}_i$ is the $tf - idf$ vector of the Wikipedia article text corresponding to concept i and cos is the cosine function which computes the similarity between the two vectors.

- $OrderRank$ is a measure which takes larger values for concepts that appear at the beginning of the document, based on the observation that important words often occur at the beginning of a document. Formally it is defined as:

$$OrderRank_{j,i} = 1 - \frac{arraypos_i}{|j|} \tag{4}$$

  where : $arraypos$ is an array containing all words of the document in the order that they occur in the document, $arraypos_i$ represents the position of the first occurrence of concept i in the array (if a concept consists of more than one word, then we take into consideration the position of occurrence of the first word of the concept) and $|j|$ is the size of document j, i.e. how many words form the document.

- $Keyphraseness$ is a global measure adapted from [7], which has a specific value for each different concept, regardless of the document we refer to, and is an indication of how much descriptive and specific to a topic a concept is. It is defined as:

$$Keyphraseness(i) = \frac{BackLinks_i}{PageHits_i} \tag{5}$$

  A concept with high $Keyphraseness$ value has more descriptive power than a concept with low $Keyphraseness$ value, even if the latter may occur more times in Wikipedia, but less times as a link. $Keyphraseness$ is normalized in the interval $[0, 1]$, after the extraction of all concepts from all documents in the corpus, so that the highest $Keyphraseness$ value is set to 1 and the lowest to 0.

After completing the disambiguation process, we linearly combine features (1) to (4) in order to construct a vector representation for each document. The final weight of concept i in document j is given by the following equation:

$$\begin{aligned} Weight(j,i) = \alpha * WFreq_{j,i} + \beta * LinkRank_{j,i} + \gamma * OrderRank_{j,i} + \\ +(1 - \alpha - \beta - \gamma) * ConceptSim_{j,i} \end{aligned} \tag{6}$$

The coefficients α, β and γ are determined by experiments and their value range is the interval $[0, 1]$.

4 Conceptual Hierarchical Clustering

Our clustering method extends the idea of frequent itemsets [5], aiming to provide a cluster description based on the Wikipedia concepts extracted from the corpus examined. Let us introduce some definitions: (a) A *global important concept* is a concept that: has a $Keyphraseness$ value greater than a specific threshold, defined as *minimum keyphraseness threshold* and appears in more than a minimum fraction of the whole document set, defined as *minimum global frequency threshold*. A *global important k-concept-set* is a set of k global important concepts that appear together in a fraction of the whole document set greater than the minimum global frequency threshold, (b) A global important concept is *cluster frequent* in a cluster C_m, if the concept is contained in some minimum fraction of documents assigned to C_m, defined as *minimum cluster support* and (c) The *cluster support* of a concept in a cluster C_m is the percentage of documents in C_m that contain this specific concept.

The method consists of two steps. At the first step, initial clusters are constructed (based on the $Keyphraseness$ of concepts and on the frequency of concepts and concept-sets using definitions (a) through (c)) where the *cluster label* of each cluster is defined by the global important concept-set that is contained in all documents assigned to the cluster. At the second step, clusters get disjoint according to a $Score$ function which shows how "good" a cluster C_m is for a document Doc_j :

$$\begin{aligned} Score(C_m \leftarrow Doc_j) = {} & [\sum_x Weight(j,x) \cdot cluster\_support(x)] \\ & -[\sum_{x'} Weight(j,x') \cdot Keyphraseness(x')] \end{aligned} \tag{7}$$

where : x represents a global important concept in Doc_j, which is cluster-frequent in C_m, x' represents a global important concept in Doc_j, which is not cluster-frequent in C_m, $Weight(j,x)$ is the weight of concept x in Doc_j as defined by Equation (6), $Weight(j,x')$ similarly as the previous one, $cluster\_support(x)$ is given by definition (c), $Keyphraseness(x')$ is given by Equation (5).

A cluster tree can be broad and deep, depending on the minimum global threshold and the $Keyphraseness$ values we define, therefore, it is likely that documents are assigned to a large number of small clusters, which leads to poor accuracy. By treating one cluster as a document (by combining all the documents in the cluster) and measure its score using the $Score$ function defined by Equation (7), we are in position to define the similarity of a cluster C_b to C_a :

$$Sim(C_a \leftarrow C_b) = \frac{Score(C_a \leftarrow Doc(C_b))}{\sum_x Weight(Doc(C_b),x) + \sum_{x'} Weight(Doc(C_b),x')} + 1 \tag{8}$$

where : $Doc(C_b)$ stands for combining all the documents in the subtree of C_b into a single document, x represents a global important concept in $Doc(C_b)$ which is also cluster frequent in C_a, x' represents a global important concept in $Doc(C_b)$ which is not cluster frequent in C_a, $Weight(Doc(C_b),x)$, $Weight(Doc(C_b),x)$

are the weights of concepts x and x' respectively in document $Doc(C_b)$. To explain the normalization by the denominator in (8), notice that, in the *Score* function, the *Cluster_Support* and *Keyphraseness* take values in the interval $[0,1]$, thus the maximum value of *Score* function would be $\sum_x Weight(j,x)$ and the minimum value $-\sum_{x'} Weight(j,x')$. So, after the proposed normalization, the value of *Sim* would be in the interval $[-1,1]$. To avoid negative values for similarity, we add the term $+1$ and we end up with the above equation. Please notice that the range of the *Sim* function is $[0,2]$.

The cluster similarity between C_a and C_b is computed as the geometric mean of the two normalized scores provided by Equation (8) :

$$Similarity(C_a \longleftrightarrow C_b) = \sqrt{Sim(C_a \leftarrow C_b) \times Sim(C_b \leftarrow C_a)} \tag{9}$$

In our method, *Similarity* value 1 is considered the threshold for considering two clusters similar. The pruning criterion computes the *Similarity* function between a child and its parent and is activated when the value of *Similarity* is larger than 1, i.e. the child is similar to its parent. Sibling merging is a process applied to similar clusters at level 1 (recall that child pruning is not applied at this level). Each time, the *Similarity* value is calculated for each pair of clusters at level 1 and the cluster pair with the highest value is merged.

5 Experiments

We evaluated our method by comparing its effectiveness with two of the most standard and accurate document clustering techniques: Hierarchical Agglomerative Clustering (HAC) (the UPGMA variant) and k-Nearest Neibghbor (k-NN) (the bisecting k-NN variant). Two well-known datasets were used for the evaluation, 10.000 documents from the 20-newsgroup collection of USENET news group articles and 6.000 documents of the Reuters 21578 dataset. For the evaluation of clustering quality we adopt a quality measure widely used in text clustering techniques, the *F-measure* [**?**].

We experimented with various values for the α, β and γ parameters of Equation (6) in order to define the effect of $WFreq$, *LinkRank*, *OrderRank* and *ConceptSim* on document representation. *LinkRank* and *ConceptSim* have the biggest effect on document representation with weights 0.4 and 0.3 respectively, whereas $Wfreq$'s weight is 0.2 and *OrderRank*'s is 0.1.

We also experimented on the *minimum keyphraseness threshold* (MinKeyph) and the *minimum global frequency* threshold (MinFreq) by choosing values which create clusters with descriptive labels. Numerous experiments showed that, if a dataset contains less than 5.000 documents, MinFreq should be set between 0.03 and 0.05, otherwise MinFreq should be set between 0.01 and 0.04. Experiments show that a value for MinKeyph around 0.5 always yields good results in different datasets, provided that there are at least a few hundreds of documents available.

The clustering results in comparison to those of HAC and k-NN, for the 20-NG and Reuters datasets are shown in Table 1.

Table 1. Experimental Results

	Dataset F-measure		Improvement	
Clustering method	20-NG	Reuters	20-NG	Reuters
HAC	0.452	0.521	**80.09%**	**58.92%**
k-NN	0.671	0.737	**21.31%**	**12.35%**
Proposed	0.814	0.828		

6 Conclusions - Future Work

In this paper, we proposed a novel method for Conceptual Hierarchical Clustering of documents using knowledge extracted from Wikipedia. The proposed method exploits Wikipedia textual content and link structure in order to create a rich and compact document representation which is built real-time using the Wikipedia API, whereas the clustering approach is hierarchical. We are currently investigating ways to improve the proposed clustering technique. These include the introduction of a novel disambiguation method, the improvement of clustering accuracy by introducing new strategies and the application of the concept based representation model to text classification tasks.

References

1. Banerjee, S., Ramanathan, K. and Gupta, A.: Clustering short texts using Wikipedia. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2007) 787–788
2. Wang, P. and Domeniconi, C.: Building Semantic Kernels for text classification using Wikipedia. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008) 713–721
3. Gabrilovich, E. and Markovitch, S.: Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In Proceedings of the 21st National Conference on Artificial Intelligence (2006) 1301–1306
4. Hu, J., Fang, L., Cao, Y., Zeng, H., Li, H., Yang, Q., and Chen, Z.: Enhancing text clustering by leveraging Wikipedia semantics. In Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (2008) 179–186
5. Fung B., Wang K., Ester M.: Hierarchical Document Clustering Using Frequent Itemsets. In Proceedings of the SIAM International Conference on Data Mining (2003)
6. Marcus, M., Santorini, B., and Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. In Computational Linguistics (1993) Volume 19, Number 2, 313–330
7. Mihalcea, R. and Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the Sixteenth ACM Conference on information and Knowledge Management (2007) 233–242
8. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of the International Conference on New Methods in Language Processing (2004)