Clustering Documents Using a Wikipedia-Based Concept Representation

Anna Huang, David Milne, Eibe Frank, and Ian H. Witten

Department of Computer Science, University of Waikato, New Zealand
{lh92,dnk2,eibe,ihw}@cs.waikato.ac.nz

Abstract. This paper shows how Wikipedia and the semantic knowledge it contains can be exploited for document clustering. We first create a concept-based document representation by mapping the terms and phrases within documents to their corresponding articles (or concepts) in Wikipedia. We also developed a similarity measure that evaluates the semantic relatedness between concept sets for two documents. We test the concept-based representation and the similarity measure on two standard text document datasets. Empirical results show that although further optimizations could be performed, our approach already improves upon related techniques.

1 Introduction

Clustering is an indispensable data mining technique, particularly for handling large-scale data. Applied to documents, it automatically groups ones with similar themes together while separating those with different topics. Creating a concise representation of a document is a fundamental problem for clustering and for many other applications that involve text documents, such as information retrieval, categorization and information extraction. Redundancy in feature space adds noise and often hurts subsequent tasks. This paper follows our previous work on using Wikipedia to create a *bag of concepts* (BOC) document representation [6]. By *concept* we mean the abstract unit of knowledge represented by a single Wikipedia article. We extend previous work by exploring the semantic relatedness between concepts to calculate the similarity between documents. In the previous work, documents are connected based on the overlap of the concepts that appear in them: this does not take account of the fact that concepts are clearly related to each other. We now explicitly incorporate the semantic connections among concepts into the document similarity measure. This allows us to identify topics that are distinct and yet relate closely to each other—*USA* and *America*, for example—and connect documents at the semantic level regardless of terminological idiosyncrasies. The experiments (Section 4) show that our BOC model together with the semantically enriched document similarity measure outperform related approaches.

Techniques such as *Latent Semantic Indexing* (LSI) [2] and *Independent Component Analysis* (ICA) [7] have been applied to the bag-of-words (BOW) model to find latent semantic word clusters. Representing documents with these clusters also allow subsequent clustering to relate documents that do not overlap in the original word space. In a quest for comparisons with our document similarity measure, we apply LSI and ICA to

T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 628–636, 2009.

our BOC model, and use the identified latent concept structures as features for clustering. Empirical results show that clustering using these latent structures is outperformed by using the plain BOC model, either with or without the enriched document similarity measure.

The paper proceeds as follows. The next section briefly describes our approach for identifying concepts in a document, each concept being associated with a Wikipedia article. Section 3 extends the semantic relatedness measure between concepts introduced in [10] to compute the semantic similarity of two documents, which forms a basis for clustering. Section 4 presents experiments and discusses results. Related work is reviewed in Section 5; Section 6 concludes the paper.

2 Representing Documents as Wikipedia Articles

In this section we describe our approach for identifying concepts in a document. There are three steps in total: identifying candidate phrases in the document and mapping them to anchor text in Wikipedia; disambiguating anchors that relate to multiple concepts; and pruning the list of concepts to filter out those that do not relate to the document's central thread. The method presented here differs from our previous approach in the way it measures the salience of each concept identified in a document and how it selects the best ones to represent the document.

2.1 Selecting Relevant Wikipedia Concepts

The first step is to map document terms to concepts in Wikipedia. Various approaches have been proposed [3,15,5]. We take the same route as [9], and use Wikipedia's vocabulary of anchor texts to connect words and phrases to Wikipedia articles. Given a plain text document, we first find phrases in it that match Wikipedia's anchor text vocabulary. For example, Wikipedia articles refer to our planet using several anchors, including *Earth*, *the world* and *the globe*. If any of these phrases appear in the document, the article about Earth will be identified as a candidate descriptor. We confine the search for phrases to individual sentences.

2.2 Resolving Ambiguous Terms

Anchors may be ambiguous in that they may refer to different concepts depending on the articles in which they are found. For example, *Pluto* links to 26 different articles, including the celestial body, the Greek god, the Disney character, and a rock band from New Zealand. Disambiguating and selecting the intended concept is essential for creating a correct thematic representation. We use machine learning to identify the correct sense. The input to the classifier is a set of possible targets for a given anchor text and the set of all unambiguous anchors from the surrounding text, which are used as context. The classifier predicts, for each sense, the probability of it being the intended one. The sense with the highest probability is selected. More details about the algorithm can be found in [9].

2.3 Pruning the Concept List

The resulting list of concepts, which together cover the topics mentioned in the input document, is rather long, because phrases are matched against a huge vocabulary; the Wikipedia snapshot we used (dated Nov. 2007) contains just under five million distinct anchors after lower casing. Irrelevant or marginally related concepts must be pruned: they add noise to the representation, which adversely impacts the document similarity calculation and reduces clustering performance. Pruning is based on salience: the average strength of relationship with the other concepts in the document. Let U denote the set of concepts extracted from a document, and salience of concept $c_i \in U$ is defined by:

$$SAL(c_i) = \frac{\sum_{c_j \in U, i \neq j} SIM(c_i, c_j)}{|U|}, \quad (1)$$

where c_j represents the other concepts in U and $|U|$ is the total number of concepts identified in the document. The more concepts c_i relates to and the greater the strength of those relationships, the more salient c_i is. The salience formula depends on $SIM(c_i, c_j)$, the semantic relatedness between two concepts. For this we use Milne and Witten's similarity measure [10]. All concepts in the list are ranked in descending order of SAL, and a fixed proportion t is discarded from the bottom of the list. In our experiments t is set to 0.1 based on empirical observations that this yields the best representation.

It is worth noting that the computational complexity of the above approach is in general linear with the input document length. The disambiguation classifier can be built beforehand, and computing the relatedness between two concepts is a linear operation. The only non-linear calculation is the last step where the averaged relatedness with all the other concepts is computed for each concept in the document. However, this step is restricted to the set of concepts identified from one document and normally the number of concepts per document is moderate. For example, the two datasets used in our experiments have, on average, 24 (OHSUMed) and 20 (Reuters) concepts per document (before pruning) respectively.

3 A Semantically Enriched Document Similarity Measure

Document similarity is typically measured using the cosine of their word vectors, so that matches indicate relatedness and mismatches indicate otherwise. Our new representation allows us to take conceptual relatedness, rather than just lexical overlap, into account. A document d_i is represented by a set of concepts U_i, each with a weight $w(c, d_i)$ (TFIDF value in our experiment). We extend the semantic relatedness between concepts mentioned earlier to the similarity between documents. Given two documents d_i and d_j, their semantic similarity is defined as:

$$Sim_{sem}(d_i, d_j) = \frac{\sum_{\forall c_k \in U_i, \forall c_l \in U_j} w(c_k, d_i) \times w(c_l, d_j) \times SIM(c_k, c_l)}{\sum_{\forall c_k \in U_i, \forall c_l \in U_j} w(c_k, d_i) \times w(c_l, d_j)}. \quad (2)$$

Because $SIM(c_k, c_l)$ is always in [0,1], Sim_{sem} is also bounded within [0,1]. 0 indicates topics in one document are completely unrelated to those in the other, and 1 indicates they are the same topics.

Table 1. Relatedness between the four concepts

	Computer Science (*CS*)	Machine Learning (*ML*)
Data Mining (*DM*)	0.45	0.80
Database (*DB*)	0.51	0.49

We then define the overall similarity between documents d_i and d_j as a linear combination of the cosine similarity Sim_{cos} and Sim_{sem} between two concept vectors:

$$DSim(d_i, d_j) = (1 - \lambda)Sim_{cos}(d_i, d_j) + \lambda Sim_{sem}(d_i, d_j). \quad (3)$$

where λ is a parameter that we set to 0.1 based on preliminary experiments.

In Hu et al.'s approach for semantically enriched document similarity [5], cosine similarity is computed on three aspects: the two document vectors, their category vectors, and their concept vectors enriched with related terms identified from Wikipedia; and the three parts are combined linearly as the final similarity measure. In our approach, the last two parts are unified neatly by a single semantic relatedness measure.

We illustrate our measure with the example used in Hu et al.'s work [5]. Given two concept sets $C_a = \{(CS, 1), (ML, 1)\}$ and $C_b = \{(DM, 1), (DB, 1)\}$, Table 1 shows the relatedness between the four concepts obtained from Milne and Witten's similarity measure. The semantic similarity between document C_a and document C_b is therefore $(0.45\times1+0.51\times1+0.80\times1+0.49\times1)/4 = 0.5625$. This value is close to that obtained by Hu et al. [5], which is 0.57. It is also worth noting that this similarity measure is not only applicable to the BOC model, but also has the potential to be extended to hybrid models where words and concepts are combined, as in [6].

4 Experiments and Results

To focus our investigation on the representation rather than the clustering method, we used the standard k-means algorithm. We created two test sets, following [5], so as to compare our results with theirs[1].

- **Reuters-21578** contains short news articles. The subset created consists of categories in the original Reuters dataset that have at least 20 and at most 200 documents. This results in 1658 documents and 30 categories in total.
- **OHSUMed** contains 23 categories and 18302 documents. Each document is the concatenation of title and abstract of a medical science paper.

4.1 Methodology

Before beginning the experiments we collected all anchor texts in the November 20, 2007 Wikipedia snapshot and lower-cased them. This produced just under five million distinct phrases linking to almost all of the two million articles in the snapshot.

[1] We would like to thank Hu et al. for sharing the OHSUMed dataset.

Documents were preprocessed by selecting only alphabetic sequences and numbers, lower-casing them, and removing concepts that appeared just once across the dataset.

Each document is represented by a vector $\overrightarrow{t_d}$ of *TFIDF* values, each element being a concept. *TFIDF* is defined as $tfidf(d,t) = tf(d,t) \times log(\frac{|D|}{df(t)})$, where t is a concept, $tf(d,t)$ is its frequency in document d, $df(t)$ is its document frequency, and $|D|$ is the total number of documents in the dataset. We set the number of clusters to the number of classes in the data. Each cluster is labeled with its dominant class. Results reported are the average of 5 runs. To compare our results with previous work, we use two evaluation measures: Purity and Inverse Purity. We also use the micro-averaged F-measure [13], weighted by class size, in a separate experiment.

4.2 Evaluation of the Semantic Document Similarity

Table 2 shows how our new document similarity measure performs in clustering on the two datasets. The other rows show the performance of Hu et al.'s algorithm and the baselines to which they compared it to: the traditional BOW, a reimplementation of Hotho et al.'s WordNet-based algorithm [4], and a system that applies Gabrilovich and Markovich's document categorization approach [3] to clustering. Our system and Hu et al.'s achieve comparable results, and are the only two approaches to provide substantial improvements over the baseline. We obtained better inverse purity because classes are more concentrated into clusters rather than dispersed across multiple clusters.

To further explore the differences between these approaches, let us take a closer look at one document that was clustered. Table 3 compares some of the concepts produced when each of the systems is asked to cluster Reuters document #15264 (results for other approaches were taken from [5]). This document discusses ongoing attempts by Teck Cominco—a Canadian mining company—to begin a joint copper-mining venture in Highland Valley, British Columbia. All of the approaches are able to pick up on the different minerals and units—*copper*, *silver*, *ounce*—and will (implicitly or explicitly) relate to synonyms such as *Cu* and *oz*. The first system, by Hotho et al., does so using WordNet, a lexical rather than encyclopedic resource. Thus it fails to pick up specific named entities such as *Teck Cominco*, but will identify terms that do not resolve to Wikipedia articles, such as *complete*. Each of the terms shown in the table can be further expanded with WordNet semantic relations; *copper* can be expanded with the associated term *cupric* and the hypernyms *metallic element*, *metal* and *conductor*.

All of the latter three approaches use Wikipedia. The approach inspired by Gabrilovich and Markovich gathers Wikipedia concepts through term overlap with

Table 2. Comparison with related work in terms of clustering purity

Dataset	Reuters			OHSUMed		
	Purity	Inverse	Impr.	Purity	Inverse	Impr.
Bag of Words	0.603	0.544	-	0.414	0.343	-
Gabrilovich and Markovich	0.605	0.548	0.33%	0.427	0.354	3.17%
Hotho et al.	0.607	0.556	0.66%	0.435	0.358	4.72%
Hu et al.	0.655	0.598	8.62%	0.459	0.388	12%
Ours	**0.678**	**0.750**	**12.4%**	**0.474**	**0.528**	**14.5%**

Table 3. Comparing features generated by different approaches

Hotho	copper; venture; highland; valley; british; columbia; affiliate; mining; negotiation; **complete**; administration; reply; silver; ounces; molybdenum
Gabri	Teck; John Townson; Cominco Arena; Allegheny Lacrosse Officials Association; **Scottish Highlands**; Productivity; Tumbler Ridge, British Columbia; Highland High School; **Economy of Manchukuo**; Silver; **Gold (color)**; **Copper (color)**;
Hu	**Tech Cominico**; British Columbia; Mining; Molybdenum; Joint Venture; Copper
Ours	Mining; Joint venture; Copper; Silver; Gold; Ore; Management; Partnership; Product (business); Ounce; Negotiation; Molybdenum; **Teck Cominco**; Vice president; Consortium; Short ton;

the document. This unfortunately allows tenuously related concepts such as *Scottish Highlands* and the *Economy of Manchukuo* to creep into the representation and cause problems. Additionally this system performs disambiguation only indirectly, which introduces more irrelevant concepts such as *Copper (color)*.

The last two systems have the tightest representation of the document, because they only contain the Wikipedia concepts that are directly discussed. Both are then able to expand out from these concepts to identify related documents regardless of textual overlap. Hu et al.'s system considers broader topics mined from the categories to which each article belongs, and associated topics mined from the links extending out from each article. Thus *Teck Cominco* is expanded with *Mining companies in Canada* and *Con Mine* in [5]. Our system, in comparison, does not need to expand concepts beforehand. Instead it can compare any two Wikipedia concepts as required, based on the relatedness measure introduced in [10]. *Teck Cominco* is essentially expanded on demand with a huge pool of possibilities, such as different mining companies (*Codelco*, *De Beers*, and about a hundred others), tools (*Drilling rig*, *Excavator*, etc.) and locations (the *Pebble Mine* in Alaska, for example). All of these new concepts are weighted with a proven relatedness measure [10], and only the concepts that are necessary to connect two related documents are ever considered.

4.3 Latent Semantic Indexing with Concepts

As an additional experiment, we apply Latent Semantic Indexing (LSI) and Independent Component Analysis (ICA) on the BOC representation (concept vectors with TFIDF values). LSI and ICA find latent structures/independent components respectively by analyzing the *concept-document* matrix. The purpose is to use the identified latent concept clusters as features for clustering and compare its effectiveness in connecting documents that do not overlap in the original concept space with using the semantically enriched document similarity measure defined in Section 3.

The only work to our knowledge so far that uses LSI with features extracted using Wikipedia is [11], where LSI is used to reduce dimensionality and Wikipedia is used to enrich text models for text categorization. Instead we use Wikipedia to extract

Table 4. Performance of LSI and ICA on BOC model on Reuters dataset

Approach	Dimensionality	Purity	Inverse Purity	FMeasure
BOC with Sim_{cos}	2186	0.667	0.750	0.561
BOC with $DSim$	2186	0.678	0.750	0.575
BOC + LSI	546	0.353	0.450	0.195
BOC + ICA	546	0.414	0.649	0.201

concepts from the input document and apply LSI/ICA directly to the BOC model that is generated. ICA has been applied to text documents in [8] and found to produce better group structures in feature space than LSI. We used the FastICA program[2] with its default settings. For a fair comparison, the number of independent components in ICA is set to the number of eigenvalues retained in LSI. The cosine measure (Sim_{cos}) is used throughout this experiment.

Table 4 shows the performance of using latent concept groups as features for clustering on the Reuters dataset. The OHSUMed dataset could not be processed because it is computationally prohibitive. The results show that the latent concept groups are not as effective as the original concepts: using cosine similarity on the BOC model (ie. based on overlaps between concept sets) still outperforms. This could be explained by the fact that ICA and LSI are applied globally and do not use any knowledge about the categories in the datasets, so the latent semantic structures that are found do not retain sufficient discriminative information to differentiate the classes [14]. Local alternatives for LSI and ICA may be better choices; but are beyond the scope of this paper.

5 Related Work

Document representation is a fundamental issue for clustering, and methods such as BOW, bags of phrases and n-grams have been widely investigated. Explicitly using external knowledge bases can assist generating concise representations of documents. Related work in this area includes Hotho et al. [4] and Recupero [12]; both use relations defined in WordNet to enrich BOW. Techniques such as Latent Semantic Indexing and Independent Component Analysis have been used to find latent semantic structures in dataset [2,8]; each structure is a linear combination of the original features (typically words). Representing documents with these latent structures can reduce the dimensionality of feature space while retaining essential semantic information, yielding significant improvement in subsequent tasks, in information retrieval [2] for example.

Despite widespread adoption for many tasks, only a limited amount of work has investigated utilizing Wikipedia as a knowledge base for document clustering [1,5,6]. Our previous work focuses on how to generate the concept-based representation for text documents and use Wikipedia to provide supervision for active learning [6]; the present paper focuses on extending the relatedness between concepts to measuring the relatedness between documents, evaluating the impact of the semantically enriched document similarity measure on clustering, and gives a more detailed analysis of the

[2] http://www.fastica.org/

concept-based document representation. The algorithm described in this paper also differs from our previous one in how it selects the best concepts to represent each document.

Our approach differs markedly from that of Hu et al. [5]. Our process for selecting and disambiguating terms to identify relevant Wikipedia concepts draws directly on previous work [9] and has been separately evaluated against manually-defined ground truth. In contrast, theirs was developed specifically for the task and has not been investigated independently. Another significant difference is the way in which the document similarity measures are calculated. They develop their own methods of measuring similarity through Wikipedia's category links and redirects, and append this to the traditional metric obtained from the BOW model. We instead start with an independently proven method of measuring relatedness between concepts [10] that takes all of Wikipedia's hyperlinks into account, and generalize this to compare documents.

6 Conclusions

This paper has presented a new approach to document clustering that extends a semantic relatedness measure defined between concepts in Wikipedia to measure document similarity. Results on two datasets prove the effectiveness of our BOC model and the enriched document similarity measure. We also investigated clustering based on a transformed feature space that encodes semantic information derived directly from the dataset, by applying LSI and ICA to the BOC model and using the latent semantic structures instead of original concepts as features for clustering, as a comparison to using the semantically enriched document similarity. Results suggest that these techniques do not improve clustering using the BOC model when performed globally.

We also observed from our earlier work [6] that BOC model can often be improved by adding further words from the document that are not represented in the BOC model, especially when the topics involved are similar. Yet in this paper we consider only concepts, albeit with an approved similarity measure. This suggests a hierarchical approach: first cluster coarsely using the BOC model, and refine clusters using hybrid models like the *Replaced* model in [6]—another interesting avenue for future work.

References

1. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering Short Texts using Wikipedia. In: Proceedings of the SIGIR, pp. 787–788. ACM, New York (2007)
2. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
3. Gabrilovich, E., Markovitch, S.: Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In: Proceedings of AAAI, pp. 1301–1306. AAAI, Menlo Park (2006)
4. Hotho, A., Staab, S., Stumme, G.: WordNet improves Text Document Clustering. In: Proceedings of SIGIR Semantic Web Workshop, pp. 541–544. ACM, New York (2003)
5. Hu, J., Fang, L., Cao, Y., Zeng, H.J., Li, H., Yang, Q., Chen, Z.: Enhancing Text Clustering by Leveraging Wikipedia Semantics. In: Proceedings of SIGIR, pp. 179–186. ACM, New York (2008)

6. Huang, A., Milne, D., Frank, E., Witten, I.H.: Clustering Documents with Active Learning using Wikipedia. In: Proceedings of ICDM, pp. 839–844. IEEE, Los Alamitos (2008)
7. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley Interscience, Hoboken (2001)
8. Kolenda, T., Hansen, L.K.: Independent Components in Text. In: Girolami, M. (ed.) Advances in Independent Component Analysis, ch. 13, pp. 235–256. Springer, Heidelberg (2000)
9. Milne, D., Witten, I.H.: Learning to Link with Wikipedia. In: Proceedings of CIKM, pp. 509–518. ACM, New York (2008)
10. Milne, D., Witten, I.H.: An Effective, Low-Cost Measure of Semantic Relatedness obtained from Wikipedia Links. In: Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI), pp. 25–30. AAAI, Menlo Park (2008)
11. Minier, Z., Bodo, Z., Csato, L.: Wikipedia-Based Kernels for Text Categorization. In: Proceedings of SYNASC, pp. 157–164. IEEE, Los Alamitos (2007)
12. Recupero, D.R.: A New Unsupervised Method for Document Clustering by Using WordNet Lexical and Conceptual Relations. Information Retrieval 10, 563–579 (2007)
13. van Rijsbergen, C.J.: Information Retrieval. Butterworths, London (1979)
14. Torkkola, K.: Discriminative Features for Document Classification. In: Proceedings of ICPR, pp. 10472–10475. IEEE, Los Alamitos (2002)
15. Wang, P., Hu, J., Zeng, H.J., Chen, L., Chen, Z.: Improving Text Classification by Using Encyclopedia Knowledge. In: Proceedings of ICDM, pp. 332–341. IEEE, Los Alamitos (2007)