Short Text Similarity with Background Knowledge

THE UNIVERSITY OF SYDNEY

Henry Petersen
Supervisor: Dr Josiah Poon

School of Information Technologies
Faculty of Engineering & Information Technologies

Motivation

Traditional methods for clustering text often represent documents using a model derived from a bag-of-terms. In such a space similarity measures are often derived from the overlap of common words between two documents.

In practice such similarity functions can be unsuitable for short text, where each document may contain only a few words. Additional information from external data sources may be required to achieve satisfactory performance.

Similar, but no common words:

"Support Vector Machine"
"SVM"

Dissimilar, but share a common word:

"Quantum Mechanics"
"Quantum of Solace"

Contributions

- Proposed a similarity measure using 'Background Knowledge' that is suitable for short text problems.
- While external data sources have previously been used for unsupervised learning, there are far fewer restrictions placed on the size and structure of the Background Knowledge collection when compared to alternative methods.
- We proposed an algorithm for automatically obtaining effective Background Knowledge and demonstrated its efficacy for unsupervised categorisation tasks.

Proposed Similarity Measure

We propose to obtain additional information from **unlabelled text relevant to the problem domain**. This unlabelled text need not be of the same length or structure as the short text being clustered; the only requirement is that it describes topics from the domain at hand. This type of external data is referred to as **Background Knowledge** and to the best of our knowledge has to date been used exclusively for supervised tasks.

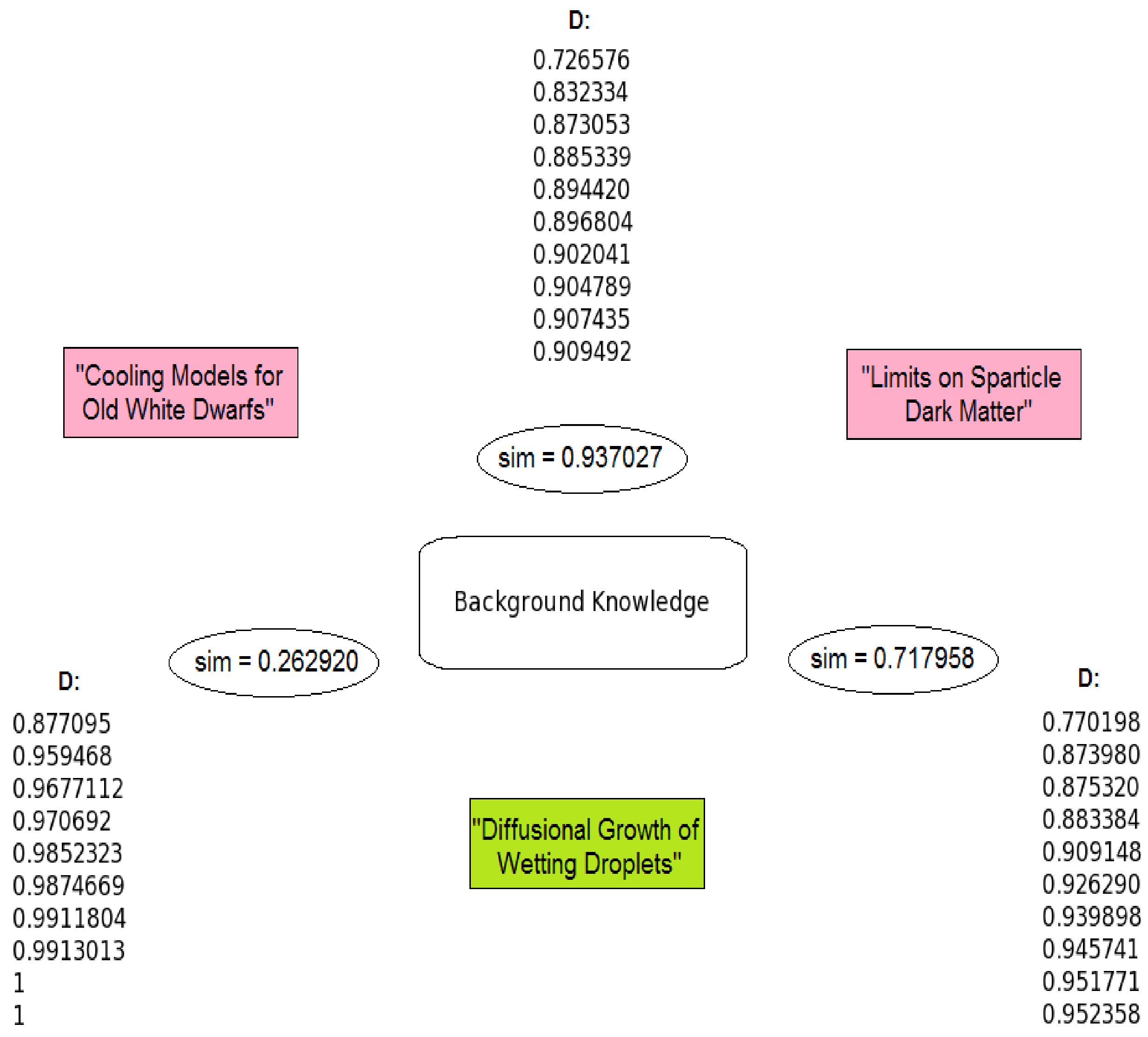
We extract from the Background Knowledge additional information on the co-occurrence structure of words in the problem domain. Observe that while two related words may not appear together in the collection of short text, they have a higher chance of doing so in the set of larger Background Knowledge.

Let BG = { b_1, b_2, ..., b_N } describe a collection of Background Knowledge, and let u_1 and u_2 denote two short text documents. Assume that all term vectors are expressed over an identical vocabulary and have been normalised to have a length of 1. A vector D of N comparison scores between u_1 and u_2 through the Background Knowledge is then created, defining each element in D as follows:

$$d_i = 1 - u_1 \cdot b_i \times u_2 \cdot b_i$$

The vector D is then sorted in increasing order and the values are normalised. The similarity between u_1 and u_2 is then defined as 1 minus the product of the smallest 30 values.

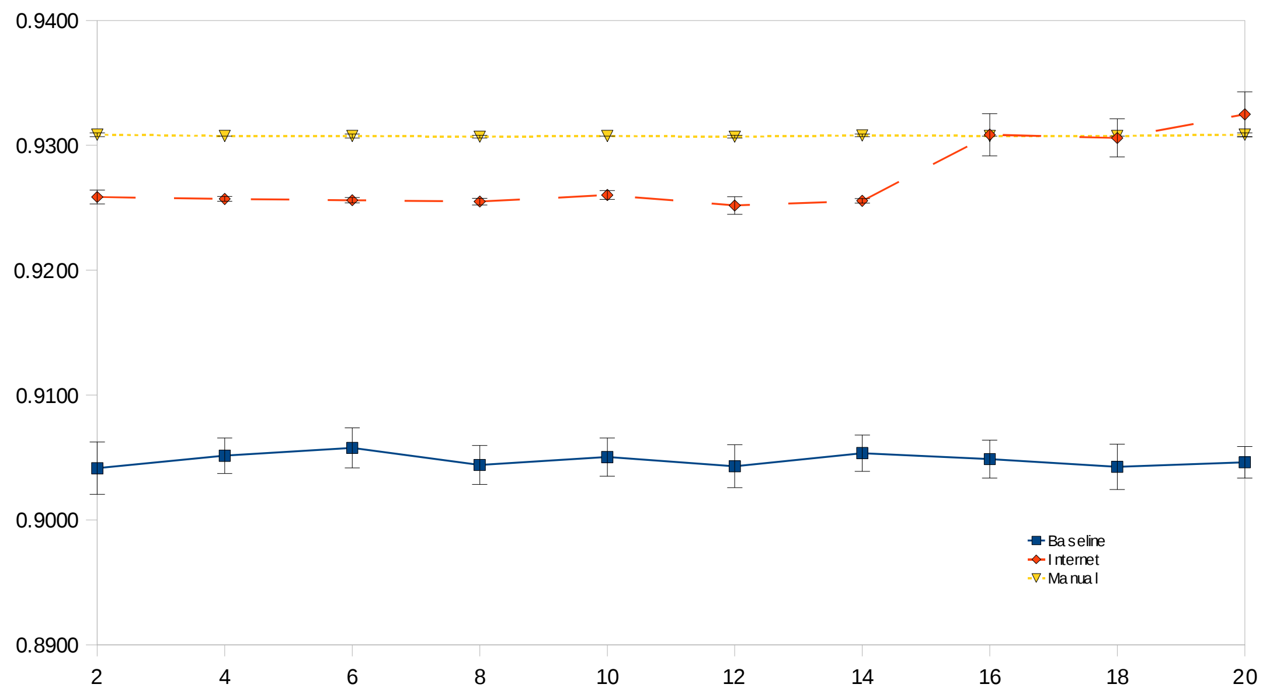
$$\text{sim}(u_1, u_2) = 1 - \prod_{i=1}^{30} \frac{d_i}{\max_{j,k}(u_1 \cdot b_j \times u_2 \cdot b_k)}$$



Example of similarity scores using the proposed measure for two astrophysics technical paper titles and one condensed matter physics technical paper title using paper abstracts as Background Knowledge

Performance

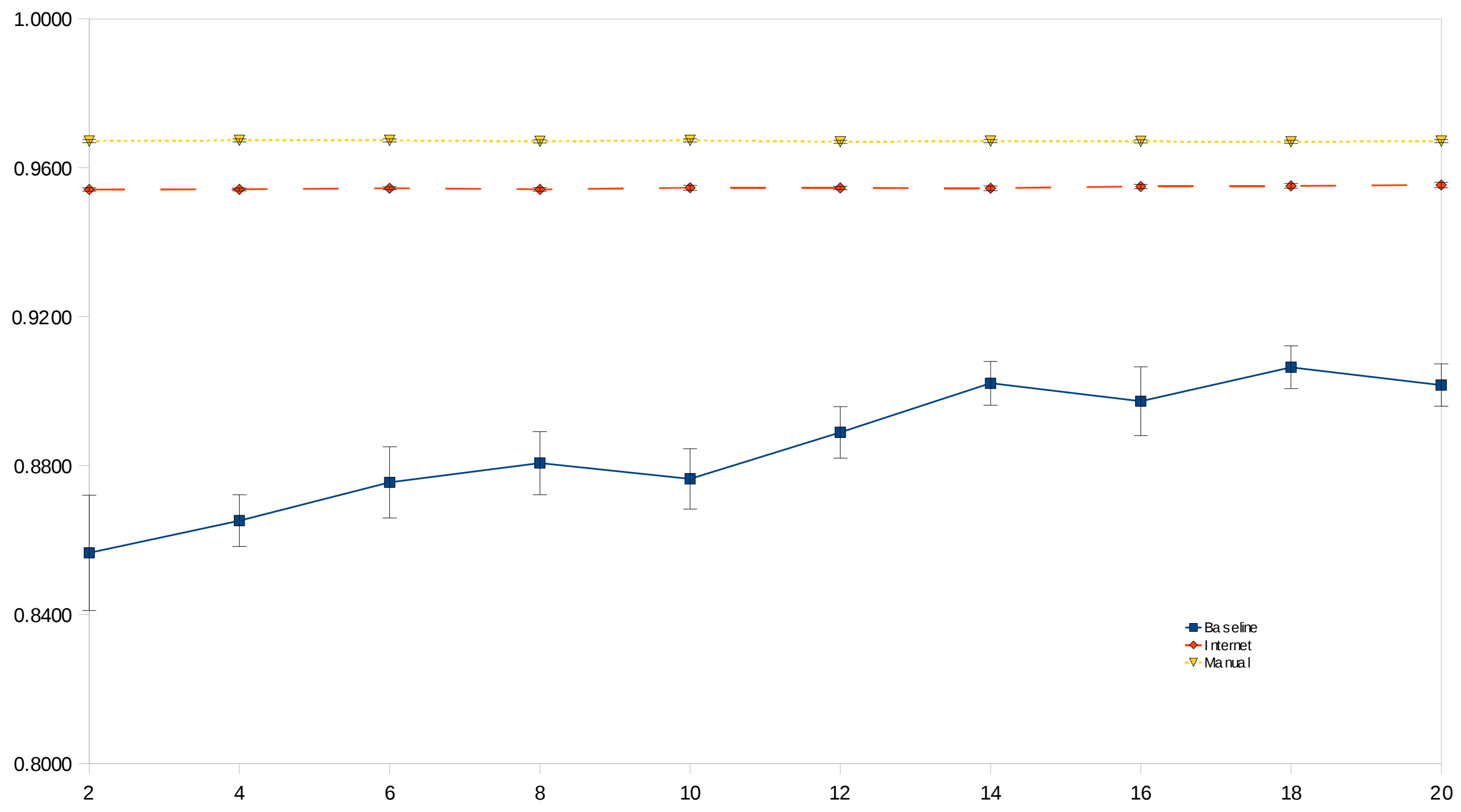
A significant increase in cluster purity can be observed when clustering short text using the proposed similarity measure as opposed to using the standard cosine similarity function. The use of Background Knowledge with the proposed function appears to increase the performance of the clustering algorithm across a range of cluster numbers and datasets.



Cluster purity with the proposed similarity function when clustering astrophysics and condensed matter physics technical paper titles

The results shown are generated using the repeated bisections algorithm from the CLUTO clustering toolkit. The graph on the left describes the purity when clustering a set of physics technical paper titles from the fields of astrophysics and condensed matter physics using paper abstracts as Background Knowledge. The graph below was obtained by clustering a set of news tiles on the topics of sport and business. Text from full articles was used as Background Knowledge.

Experiments demonstrate that both the automatically and manually obtained Background Knowledge can provide significant additional information to the clustering algorithm. In general the Background Knowledge generated automatically by the proposed algorithm is slightly outperformed by manually chosen collections, which indicates room for improvement in our method.



Cluster purity with the proposed similarity function when clustering news article titles from the topics of sports and business

Automatically Generating Background Knowledge

We demonstrate that effective Background Knowledge for unsupervised problems can be **obtained automatically from internet search engines** such as Google.

Literature has shown that **HTML pages** can be used as Background Knowledge. Search engine queries to obtain these pages previously have been generated using terms with the greatest information gain for each class, however these methods only work for supervised tasks.

Latent Semantic Analysis (LSA) is a technique that automatically discovers and ranks the importance of latent concepts (or classes) described by a collection of text. LSA describes concepts as linear combinations of terms, which can be used to rank words according to their relevance for each latent concept in a text collection.

ray	quantum	galaxies
gamma	spin	ray
observations	model	galaxy
galaxy	dimensional	infrared
galaxies	phase	spiral

Top 5 terms for the three main concepts extracted from a set of astrophysics and condensed mater physics technical paper titles

- Consider the 10 top ranked concepts
- Get 10 top ranking words for each concept
- Create queries using 3 word combinations of words from each concept
- 1200 queries in total
- Collect up to 10 HTML pages from the search engine for each query