Clustering of Medical Publications for Evidence Based Medicine Summarisation

Sara Faisal Shash and Diego Mollá

Department of Computing,
Macquarie University, Sydney, 2109 NSW, Australia
{sara-faisal.shash,diego.molla-aliod}@mq.edu.au
http://web.science.mq.edu.au/~diego/medicalnlp/

Abstract. We present a study of the clustering properties of medical publications for the aim of Evidence Based Medicine summarisation. Given a dataset of documents that have been manually assigned to groups related to clinical answers, we apply K-Means clustering and verify that the documents can be clustered reasonably well. We advance the implications of such clustering for natural language processing tasks in Evidence Based Medicine.

1 Introduction

Evidence Based Medicine (EBM) is the practice that highlights the use of proven and current medical research and literature, when making clinical decisions. The process of EBM requires physicians to search, read and appraise medical literature in order to obtain recommendations for decisions. However, research has shown that accurate evidence in EBM is retrieved using a time consuming and resource intensive process that is largely manual and does not take advantage of emerging information processing technologies [1].

This paper contributes to solve this problem by outlining the application of document clustering to help identify the clusters of documents relevant to a given question. This will contribute to the eventual construction of an evidence based summary and create clusters of reference documents that will ultimately allow medical practitioners to improve their effective practice of EBM.

2 Clustering for Evidence Based Medicine

The ultimate goal of our research is to build a query-based multi-document summarisation system that takes, as input, a clinical question and a list of relevant documents, and generates a summary of the key relevant information extracted from the original documents that is relevant to the clinical question.

Document clustering is an unsupervised machine learning task that aims to discover natural groupings of data [2]. Document clustering has been used to aid the practice of EBM in various ways. Work done by Pratt and Fagan [3] showed that organising medical search results into meaningful groups that correspond to

a given query increases the efficiency of the search experience for users. Lin and Demner-Fushman [4] also show how grouping MEDLINE citations into clusters, based on extracted interventions from document abstract texts, improves the understanding of literature search results. A text mining framework for assisting bio-medical researchers through automatic document clustering and ranking was also developed by Lin et al. [5].

For the present study we use a corpus of clinical questions and evidence-based summaries obtained from the "Clinical Inquiries" section of the Journal of Family Practice (JFP) [1] [6]. The corpus is freely available and comprises 456 questions. Each question is accompanied with the group of documents from which answers are obtained. The answer to a clinical question in the corpus has several parts. Each part has a number of documents associated to it.

It is our goal to determine whether traditional clustering techniques applied to the set of documents relevant to a clinical question can be used to re-create the groups of documents relevant to the answer parts. Thus, we will perform 456 distinct clustering tasks and compare the resulting clusters with the document groupings in our dataset. In this paper, we will name the clusters produced by our method "clusters", and the clusters defined in the annotated data "source clusters". It is anticipated that the clustering criteria used in each question will be different, and therefore separate clustering methods would be required. In this paper, however, we study effect of clustering techniques without using the question information, as a first step towards query-based summarisation.

3 Clustering Experiments

In the original data set, documents were in the PubMed XML format[2] that comprises the article's abstract and metadata such as the title of the article, publication type, author, year of publication, medical subject headings (MeSH),[3] and country. We used MetaMap [7], a program developed to map biomedical terms to concepts in the Unified Medical Language System (UMLS), to select the medical terms from the text. We then conducted preliminary clustering experiments on four representations of the data set: (i) whole XML (original format), (ii) abstracts of articles only, (iii) terms that have an UMLS concept, and (iv) UMLS medical semantic types. Words were lowercased, stopwords removed, and remaining words were weighted based on *tf.idf*.

We used K-means as the clustering approach, using the original numbers of source clusters as the K parameter. This value of K is different for each question. Since this assumes prior knowledge we can take the result as an upper bound.

To determine clustering quality we used the cluster entropy measure. The entropy of cluster i is:

[1] `http://jfponline.com`
[2] `http://www.nlm.nih.gov/bsd/licensee/data_elements_doc.html`
[3] `http://www.nlm.nih.gov/mesh/`

$$Entropy(i) = -\sum_{j} p_{i,j} \log_2 p_{i,j}$$

Where $p_{i,j}$ is the number of documents in cluster i that belong to source cluster j, divided by the number of documents in cluster i.

The entropy measure of the clusters generated in a particular question of our data set is the weighted average of the entropies of all clusters from the question, where the weight is a ratio of the cluster size relative to the total set of documents relevant to the question. We then computed the average entropy across all questions. The results are in Table 1.

Measure	Whole XML	Abstract only	Concepts only	Semantic types
Euclidean	0.260	0.264	0.274	0.310
Correlation	0.348	0.362	0.349	0.347
Cosine	**0.249**	0.266	0.277	0.298
Dice	0.332	0.328	0.324	0.334
Jaccard	0.320	0.330	0.317	0.327
Manhattan	0.288	0.299	0.305	0.296

Table 1. Average entropy for optimal K clusters. The best result is marked in bold.

To interpret the results, note that purely random clustering would give an entropy of $-\log_2(1/K)$. For the average number of clusters in the dataset $K = 2.4$, the resulting entropy would be 1.263. As we can see from Table 1, the resulting clusters have much lower entropy values, indicating good clustering results. We can also observe that the lowest entropy value is obtained when Cosine Distance is used to cluster documents that are represented as whole XML documents. It is important to note that K-means clustering provides disjoint clusters that provide no provisions for clusters that overlap. At this stage, every document is assigned a unique cluster.

In regards to the representation of the data set (Whole XML, Abstract Only, Concept Only, Semantic Type Only), we can observe from Table 1 that there is little disparity between the entropy values obtained from the different representations of the data set. Entropy values for documents represented as Whole XML are, however, producing the best results (lowest entropy) in general. This might be due to the similarity between documents being able to be computed on more information, which in-turn yielded better clustering results. Entropy based on semantic types are the worst, presumably because the semantic types are too general and many words were grouped to the same semantic type. It was interesting to observe that the UMLS concepts did not produce better results than the abstracts only.

To determine the optimal number of clusters we tried the following three methods:

User defined K: This is a constant value of K for each question. We experimented with values of $K = 2$, 3, and 4 which are constant across all questions.

Rule of Thumb: Based on the total number m of documents in a cluster [8]. This provides a value of K that is distinct for each question.

$$K = \sqrt{m/2}$$

Cover Coefficient: Distance to each question and based on the number m of documents, the number n of terms, and the number t of non-zero entries in the matrix of bags of words [9],

$$K = \frac{m \times n}{t}$$

Table 2 shows the values of entropy for all our experiments. We only show the results on full XML documents since these performed best in our previous experiments.

Measure	$K = 2$	$K = 3$	$K = 4$	RoT	Cover	Original
Euclidean	0.489	0.309	0.205	0.163	0.235	0.260
Correlation	0.604	0.413	0.283	0.238	0.316	0.348
Cosine	0.479	0.298	0.213	0.154	0.224	0.249
Dice	0.572	0.368	0.250	0.204	0.290	0.332
Jaccard	0.562	0.360	0.252	0.191	0.293	0.320
Manhattan	0.522	0.327	0.226	0.174	0.281	0.288

Table 2. Average entropy for different cluster numbers.

To interpret the above results, note that the entropy values will improve (decrease) as we increase the number of clusters. Therefore one can only compare methods that use (approximately) the same number of clusters. The average number of clusters in the original setting (when the number of clusters is provided) is 2.4. The average numbers of clusters of the rule of thumb and the cover method are 3.8 and 2.8, respectively. The cover method approximates the original number of clusters, and the entropy values are second to the rule of thumb. Thus, the cover method is the best compromise for the number of clusters and the resulting entropy values. This might be because the cover method uses information specific about the words in each document and assigns a higher weighting to documents that have a lot of terms in common.

4 Conclusions and Further Work

We have studied the effect of using K-means clustering for Evidence Based Medicine (EBM). Our system attempts to reproduce the original groupings of documents that provide the clinical evidence to the different components of the

answer to a clinical question. The good entropy results demonstrated that K-Means works well in capturing these groupings.

By providing such clusterings we allow the separation of the different components of an EBM answer. This information can be used in future systems to provide the final EBM answers by applying information extraction techniques and redundancy-based approaches on the clusters.

In further work we will explore the use of alternative clustering methods such as Agglomerative Clustering or clustering based on Topic Modelling. More interestingly, we will explore the possibility of using fuzzy clustering methods that will enable a document to be assigned to multiple clusters. This will provide a closer approximation to the real scenario.

The present study considered relevance of the clusters to the question only implicitly in the sense that all documents were relevant to the question to start with. We will study the possibility of integrating supervised clustering techniques and incorporating similarity measures and clustering techniques that tightly incorporate the information of the question.

Further on the ultimate goal to produce the final EBM summaries, we will investigate methods to generate expressions of the cluster topics as means to extract the evidence relevant to the clinical answers.

References

1. Cohen, A.M., Adams, C.E., Davis, J.M., Yu, C., Yu, P.S., Meng, W., Duggan, L., McDonagh, M., Smalheiser, N.R.: Evidence Based Medicine: the essential role of systematic retrieval and the need for automated text mining tools. In: Proceedings of the 1st ACM International Health Informatics Symposium. pp. 376–380 (2010)
2. Andrews, N.O., Fox, E.A.: Recent Developments in Document Clustering. Tech. rep., Virginia Tech (2007)
3. Pratt, W., Fagan, L.: The Usefulness of Dynamically Categorizing Search Results. Journal of the American Medical Informatics Association 7(6), 605–617 (2000)
4. Lin, J.J., Demner-Fushman, D.: Semantic clustering of answers to clinical questions. In: AMIA Annual Symposium Proceedings (2007)
5. Lin, Y., Li, W., Chen, K., Liu, Y.: A Document Clustering and Ranking System for Exploring {MEDLINE} Citations. Journal of the American Medical Informatics Association 14(5), 651–661 (2007)
6. Mollá, D., Santiago-Martínez, M.E.: Development of a Corpus for Evidence Based Medicine Summarisation. In: Proceedings of the Australasian Language Technology Workshop (2011)
7. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the 2001 AMIA Annual Symposium. pp. 17–21 (Jan 2001)
8. Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Academic Press, London (1979)
9. Can, F., Esen A. Ozkarahan: Concepts and Effectiveness of the Cover-Coefficient-Based Clustering Methodology for Text Databases. ACM Transactions on Database Systems 15(4), 483–517 (1990)