Journal of Biomedical Informatics 55 (2015) 116–123

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Multi-focus cluster labeling

Line Eikvil [a], Tor-Kristian Jenssen [b], Marit Holden [a,*]

[a] Norwegian Computing Center, P.O. Box 114 Blindern, NO-0314 Oslo, Norway
[b] PubGene AS, Sognsveien 70A, PO Box 37 Vinderen, 0319 Oslo, Norway

ARTICLE INFO

Article history:
Received 18 December 2014
Revised 19 March 2015
Accepted 30 March 2015
Available online 11 April 2015

Keywords:
Text mining
Cluster labeling
Multi focus

ABSTRACT

Document collections resulting from searches in the biomedical literature, for instance, in PubMed, are often so large that some organization of the returned information is necessary. Clustering is an efficient tool for organizing search results. To help the user to decide how to continue the search for relevant documents, the content of each cluster can be characterized by a set of representative keywords or cluster labels. As different users may have different interests, it can be desirable with solutions that make it possible to produce labels from a selection of different topical categories. We therefore introduce the concept of multi-focus cluster labeling to give users the possibility to get an overview of the contents through labels from multiple viewpoints.

The concept for multi-focus cluster labeling has been established and has been demonstrated on three different document collections. We illustrate that multi-focus visualizations can give an overview of clusters along axes that general labels are not able to convey. The approach is generic and should be applicable to any biomedical (or other) domain with any selection of foci where appropriate focus vocabularies can be established. A user evaluation also indicates that such a multi-focus concept is useful.

1. Introduction

Technologies enabling information retrieval, like indexing and searching, have become very important as the amount of information that is digitally available continues to grow exponentially. As the total amount of available information grows, so will the amount of information returned by a query. Hence, organization of the returned information will become just as important as the retrieval.

Topical organization of document collections is an important basis for further description of the contents through e.g. labels, summaries or visual presentations. Clustering is emerging as an efficient tool for organizing search results and resolving the often inherent ambiguities of user queries, as different contexts often are naturally separated in different clusters.

When a document collection has been divided into groups of documents, a user will need some information about the different clusters to be able to decide how to continue the search for the most relevant documents. A challenge is then to present the clusters to the user in a way that gives an overview of the contents. A list of the documents in each cluster, sorted with respect to some criterion, is a common way of presenting the results. However, as each cluster might contain hundreds of documents, other more informative ways of summarizing the information are needed. A simple and intuitive way of doing this is to assign a set of representative keywords or labels to each cluster. However, studies of manual label assignment show that the choice of labels is subjective and will be dependent on each person's judgment, preferences and interests [1]. As a solution to this we therefore introduce the concept of *multi-focus cluster labeling* giving users the possibility to get an overview of the contents through labels from multiple viewpoints. This can also provide views into the document collection along other axes than the clustering does, giving multiple views into the same set without re-clustering.

We will focus on document collections of biomedical papers. In [2] a list of user needs has been collected from previous studies on needs of biomedical specialists. A keyword here is simplicity, where users prefer to use short and simple queries, want a familiar and simple interface, are sensitive to information overload, and when presented with ranked lists few users review results beyond the first page. Hence, an additional aim of our multi-focus approach is that it should be simple and intuitive for the users.

2. Related work

The aim of our work is to give the user a better overview of the documents returned by a specific search in a biomedical document repository. Such searches will generally return a large number of

* Corresponding author.
E-mail addresses: line.eikvil@nr.no (L. Eikvil), tkj@pubgene.com (T.-K. Jenssen), marit.holden@nr.no (M. Holden).

http://dx.doi.org/10.1016/j.jbi.2015.03.012
1532-0464/

documents, and various approaches have been suggested for trying to help the user getting a better overview.

One approach to achieve better overview of returned search results is to cluster the documents, and then trying to describe these clusters [3]. Different approaches for describing the clusters or other collections of documents have been studied, e.g. through keywords [3], textual summaries [4], related topics [5], topic assignment [6] or through relationships e.g. between documents [7], terms [8] or authors [9].

In general these approaches offer one way of seeing the clusters. However, we see that when cluster labels are assigned manually, the labels will differ from person to person and often be dependent on each person's judgment, preferences and interests. In our approach we will therefore try to cater for this by offering multiple views into the clusters, through a multi-focus approach.

There are a few other studies where different views into retrieved document sets have been proposed. Anne O'Tate [10] provides clustering by topic, but also grouping by journals, author, year and affiliations. GoPubMed [11] offers clustering by MeSH terms (Medical Subject Headings) or GO-terms (Gene Ontology), and grouping by author, location and journal. The BioPrompt-Box [12] offers different groupings of documents returned as query results, where the user can choose between different properties for the clustering such as keywords, organism names and GO-terms. These approaches give the user different views into the returned search results, but all the options will result in a different grouping of the results. Our approach will instead give *different views* into the *same clusters*.

Our aim is to achieve this through an approach that is generic, fast and simple to use, and present the results in a way that makes them easy to interpret by using a compact visualization offering overview at-a-glance.

3. Methods

In our approach the search results are organized into clusters and then an overview of the contents of these clusters is provided by offering different views into document clusters through our concept of multi-focus labeling. In the following we describe the methods used for clustering, labeling and visualization and define our multi-focus concept.

3.1. Clustering

Document clustering aims to partition an unlabeled sample set of documents into a predefined number of disjoint clusters through an unsupervised, exploratory process.

Clustering methods are either based on agglomerative algorithms (e.g. hierarchical clustering) or on partitional algorithms (e.g. *k*-means clustering). Partitional algorithms have in experiments [13,14] been shown to lead to better clustering solutions than agglomerative algorithms, which suggests that partitional clustering algorithms are well-suited for clustering large document datasets due to not only their relatively low computational requirements, but also comparable or even better clustering performance.

Based on this we have chosen to use *k*-means clustering for our sets of biomedical abstracts. Our multi-focus labeling is however independent of the approach used for clustering and may be combined with any method.

3.2. Cluster labeling

Manning et al. [15] divide cluster labeling approaches into cluster-internal and differential approaches. Cluster-internal methods are efficient, but they fail to distinguish terms that are frequent in the collection as a whole from those that are frequent only in the cluster. Differential cluster labeling selects cluster labels by comparing the distribution of terms in one cluster with that of other clusters.

For user interfaces where humans interact with clusters, it is crucial to label the clusters so that the users can see what a cluster is about. Still, Manning et al. [15] point out that comparatively little work has been done on labeling clusters. One method based on a differential cluster labeling scheme, suggested by Popescul and Ungar [16], has however obtained good results. This is also one of very few algorithms that are independent of the clustering technique used. We will therefore base our cluster labeling on this method.

This differential cluster labeling scheme selects cluster labels by comparing the distribution of terms in one cluster with that of other clusters' frequency, and is based on computed frequency ($p(w|cl)$) and predictiveness ($p(w|cl)/p(w)$):

$$\text{score}(w|cl) = p(w|cl) \cdot \frac{p(w|cl)}{p(w)}, \tag{1}$$

where $p(w|cl)$ is the local probability of a word w, i.e. the probability of a word w in the cluster cl, and $p(w)$ is the global probability of a word, i.e. the probability of a word w in the entire document collection. This means that the Popescul and Ungars method for cluster labeling gives high weights to words that occur often in a cluster (frequency), and at the same time separates this cluster from the other clusters (predictiveness).

3.3. Multi-focus cluster labeling

Accurate and comprehensible cluster labels let the user comprehend the collection's content faster [17]. Hence, the effectiveness of a cluster label can be said to be related to a user's success in identifying relevant clusters and documents. However, for the same set of documents, different users may have different objectives and thereby different preferences in terms of labels.

The idea of multi-focus cluster labeling is therefore to present a user with several sets of cluster labels, one for each focus of interest for the user. Furthermore, the aim is to enable such multi-focus views without re-clustering or heavy computations.

3.3.1. Focus definition

In this framework, a focus can be considered as a particular aspect of a more general topic, where the topic is represented by a set of documents. As an example consider a document set on a specific disease, where symptoms and treatments represent two different aspects or foci of this disease and where documents may contain both these aspects. We assume that these aspects can be represented by a set of words and define a focus in terms of a vocabulary.

We will then need to establish focus vocabularies. Such vocabularies may be based on existing ontology or vocabulary resources. When existing vocabularies do not provide sufficient coverage, the use of automatic vocabulary expansion can be used (e.g. [18,19]). Topic models [20] can also be used to put words with similar semantics into the same group. However, completely automatic approaches for vocabulary construction are less relevant in this context where we want to have control of the definition of the focus.

In this study the vocabularies have been extracted from ontology information available in different internet resources. Details on the specific vocabularies are given in Section 4.2.