Search result clustering based on clustering context

extended abstract

Michal Meina\(^1\) and Hung Son Nguyen\(^2\)

\(^1\) Nicolaus Copernicus University, Faculty of Mathematics and Computer Science, Toruń, Poland
mich@mat.umk.pl

\(^2\) University of Warsaw, Faculty of Mathematics, Informatics and Mechanics, Warsaw, Poland
son@mimuw.edu.pl

Abstract. This paper introduces a novel, interactive and exploratory, approach to information retrieval (search engines) based on clustering. Presented method allows users to change the clustering structure by applying a free-text clustering context query that is treated as a criterion for document-to-cluster allocation. Exploration mechanisms are also delivered by redefining the interaction scenario in which the user can interact with data on the level of topic discovery or cluster labeling. In this paper, the presented idea is realized by a graph structure called the Query-Summarize Graph. This data structure is useful in the definition of the similarity measure between the snippets as well as in the snippet clustering algorithm. The experiments on real-world data are showing that the proposed solution has many interesting properties and can be an alternative approach to interactive information retrieval.

1 Introduction

Early experiments on Information Retrieval Systems such as [1][3] indicates low scalability of such systems - user studies shows that with the increase of document collection the recall of retrieved document is lowering. The main problem still is the lack of methodology in search engines that will capture and support cognitive capabilities of users. Semantic gap in the field of information retrieval can be expressed as the question: how can we transfer user demands and actions (information need) into computational model of documents and retrieval scenario?

Promising direction of research that address this semantic gap is Exploratory Search which can be seen as specialization of retrieval task where the search problem is poorly defined and the user have no or very low knowledge about domain [2]. IR systems needs to introduce new methodologies that move the search process beyond predictable fact retrieval eventually supporting Knowledge Discovery.
In this paper we will introduce complete information retrieval system that is focused on interaction with the user and defines new concept of querying for clusters. The informal definition of *clustering context* will be postulated and methodology of interaction with the user will be covered.

2 Document model

In order to maintain fine-grained control over such small feature set as it is in search result we will exploit term co-occurrence graph. Co-occurrence graph that represents document collection \( D \) is a undirected graph \( G(V, E) \) whose vertices corresponds to all possible terms \( (t \in T) \) in document space and edges between nodes are created according to the occurrence frequency of the both terms in one document. Edge between terms \( t \) and \( w \) exists only if their co-occurrence in non-zero and the weight is assigned to relate to this number. We will consider weights based on the relevancy between terms, denoted by similarity function \( sim \) (that could be obtained using some statistical document model, eg. LSI).

Query-Summarize Graph is \( G_q \) is a subgraph of \( G \) induced by a context query \( q \). This structure is used in finding cluster descriptions (topics) and then in assigning documents to those description. The semantic topic is recovered from graph using so-called “modularity” property of graph. Informally we can say that we do want maintain high cluster internal quality measures and additionally output clusters that are “semantically close” to *query context*. The idea is illustrated on Fig. 1. Detailed description on constructing such graph by greedy algorithm, finding modules and assigning document to descriptions is covered in full paper.

![Fig. 1. Searching by clustering queries](image-url)
The processing scheme consists of following steps:

**Step 1**: Selection of document collection;
**Step 2**: Definition of clustering context query \( q_c \);
**Step 3**: Discovery of topics with respect to \( q_c \);
**Step 4**: Finding a cluster descriptions (topics);
**Step 5**: Matching documents to cluster descriptions;

3 Retrieval Scenario

Knowledge Discovery is supported by interactive reorganization of document set into different clusters each time revealing different interconnections between documents explained by cluster labels. In such scenario search is understood as a process of consecutive interaction, each composed of: (1) querying (2) processing or interpreting the search result by user and (3) refining the query. Presenting the sub-result in the form of clusters of meaningful snippets is crucial element of query refinement. We can describe two main goals in interaction of the user with the data:

**Defining the context** of clustering for tightly coupled documents which is conceptually similar to applying the weighting schemata into indiscriminative feature set. In the case where document set partitioning is ambiguous user can explicitly define a feature subset (a context). The context is considered as set of topics implicated by the query.

**Context narrowing** means sequential querying against search result in which user is able to construct final context and corresponding clustering, each time narrowing the semantic field of the query.

4 Experiment

Our experimental study was twofold: (1) examination of structure of Q-S Graphs and verifying their properties regarding proposed constrains and (2) cluster quality validation.

**Data** We gathered more than half million news articles from various sources using Newsblaster\(^3\) - a newswire aggregator which was used as index for crawling. Plain text was extracted from html files and then stop list and stemming was applied. As semantic similarity we have used cosine similarity, e.g. \( \text{sim}(t, w) = \frac{t \cdot w}{|t||w|} \), where \( t, v \) are terms vectors in tf-idf weighted document-term matrix. Search result for every run of experiment was collected using lucene search engine by keyword query randomly extracted from document titles (with wasn't later used in clustering). Then another query was chosen from search result documents titles and used as clustering context query.

**Cluster quality measure** In order to investigate the topical structure of final clustering we measure homogeneity of output clusters using root-mean-square standard deviation (RMSTD), that measures homogeneity of clusters.

\(^3\) [http://newsblaster.cs.columbia.edu/](http://newsblaster.cs.columbia.edu/)
RMSTD value for every test run was scaled by baselines values of this measures from K-Means clustering and random document partitioning. Fig. 2 shows distribution of RMSTD values by query idf value. For queries with 1.75 idf value (most meaningful queries) average result was almost exactly in between baseline measures.

References