ABSTRACT

This chapter presents a new measure of document similarity – the GNGrank that was inspired by the popular opinion that links between the documents reflect similar content. The idea was to create a rank measure based on the well known PageRank algorithm which exploits the document similarity to insert links between the documents. A comparative study of various link- and content-based similarity measures, and GNGrank is performed in the context of identification of a typical document of a collection. The study suggests that each group of them measures something different, a different aspect of document space, and hence the respective degrees of typicality do not correlate. This may be an indication that for different purposes different documents may be important. A deeper study of this phenomenon is our future research goal.

DOI: 10.4018/978-1-60960-102-7.ch001
INTRODUCTION

The usual way to get an overview of a large collection of objects (e.g. documents) is to cluster them, and then to look for the representatives (or summaries) of the individual clusters. The objects are placed within a feature space (in case of documents, frequently features are the terms, and co-ordinates are e.g. tf-idf measure and the representatives are the centroids (or medoids) of clusters (Manning, Raghavan & Schütze, 2009).

There are some conceptual problems with such an approach in case of document collections. On one hand, one insists that text understanding is essential for proper clustering. Regrettably, application of the full-fledged text understanding methods for very large collections is not feasible so that some replacements have to be sought, therefore, in fact, the feature space approach is the dominant one.

The next problem is that with feature space approach a rigid weighting of features is imposed, whereas the natural language experience is that within the given group of related documents the meaning and so the importance of terms may drift. We proposed here a solution called contextual processing, where terms are re-weighted at stages of the clustering process (Ciesielski & Klopotek, 2006).

Then we have the issue of cluster relationships. Clusters formed are usually not independent. Hierarchical clustering surely does not cover all the possible kinds of relationships among clusters. For this reason we pledged for using competitive clustering methods like WebSOM, i.e. text document version of self organizing maps, (Kohonen, Kaski, Somervuo, Lagus, Oja, & Paatero, 2003), Growing Neural Gas, or GNG, of Fritzke (1997) or aiNet (an immunological method mimicking the idiotypical network) of de Castro & Timmis (2002)).

Finally, there is a problem of the centroid. The centroids are usually “averaged” documents, i.e. they represent a rather abstract concept. Averaged weights of documents may in fact not represent any meaningful document at all, and closeness to the centroid may say nothing about the importance of a document for the collection. Therefore, in our system we aim at a more realistic representative of a cluster. In this chapter we want to investigate two competing technologies:

- a histogram-based notion of document typicality
- a PageRank-like selection of “medoidal” documents.

The abovementioned concepts have been implemented and tested within our map-based search engine BEATCA (Klopotek, Wierzchon, Ciesielski, Draminski & Czerski, 2007).

Subsequently we will explain these ideas in some extent. In particular within the chapter we will explain in detail the idea of contextual clustering, methodology behind identifying typical documents and medoidal documents and show results of empirical evaluation of relationships between traditional centroids, typical documents and medoidal documents.

BACKGROUND AND MOTIVATION

The Idea of Contexts

We have introduced the idea of contexts in the paper (Ciesielski & Klopotek, 2007).

For purposes of document retrieval we rarely encounter deep linguistic analysis. For massive collection, shallow analysis is applied at most, and term space representation is most frequently encountered. In this representation, instead of analyzing the text linguistically, distinct terms are identified and frequencies of their occurrence are counted. Then a space is imagined in which all distinct terms from all documents of the collection serve as dimension labels, and documents are treated as vectors in this space where the coordinate
Related Content

A New Image Distortion Measure Based on Natural Scene Statistics Modeling
[www.igi-global.com/article/new-image-distortion-measure-based/68001?camid=4v1a](www.igi-global.com/article/new-image-distortion-measure-based/68001?camid=4v1a)

Insilico Approach for Epitope Prediction toward Novel Vaccine Delivery System Design
[www.igi-global.com/chapter/insilico-approach-for-epitope-prediction-toward-novel-vaccine-delivery-system-design/170222?camid=4v1a](www.igi-global.com/chapter/insilico-approach-for-epitope-prediction-toward-novel-vaccine-delivery-system-design/170222?camid=4v1a)

Precision, Flexibility, and Tool Support: Essential Elements of Pattern Formalization
[www.igi-global.com/chapter/precision-flexibility-tool-support/8161?camid=4v1a](www.igi-global.com/chapter/precision-flexibility-tool-support/8161?camid=4v1a)

A Highly Efficient Content Based Approach to Filter Pornography Websites
[www.igi-global.com/article/highly-efficient-content-based-approach/68005?camid=4v1a](www.igi-global.com/article/highly-efficient-content-based-approach/68005?camid=4v1a)