BI Lic 1 MAPINTEL/NOVAIMS/DSAIPA/DS/0116/2019

Project Abstract:

Competitive Intelligence (CI) is concerned with gathering and analyzing information on any aspect of the business environment (competition, customers, legal framework, etc.) needed to support executives in strategic planning and decision-making. AICEP's mission is to promote Portuguese exports and to secure foreign direct investment, playing a major role in the economic development and job creation in Portugal. CI is a key component of AICEP's activity, which requires keeping track of current affairs and sift through the endless flow of news about markets, trade, industries, countries and politics.

The ability to quickly identify, process and understand the impact of political, economic and social events in the world arena constitutes a fundamental requirement to accomplish AICEP's mission. These events can create new opportunities but can also present new threats and challenges to the Portuguese economy. Most relevant data for this purpose is unstructured, more specifically text data, in the form of news, reports, trade agreements, company announcements, etc. However, these data are not yet widely used by AICEP as a source to gather competitive strategic intelligence. The main reason behind this is the difficulty in the extraction of information from text data. In an activity in which timing and assertive responses are of the essence, having analysts sorting through the Internet to find relevant news and updates is ineffective, timeconsuming and expensive. In order to keep up with competing Nations, looking to take advantage of the same trading and investment opportunities, Portugal, through AICEP, needs to have available a modern CI infrastructure.

Document clustering can be seen as the unsupervised learning task that extracts otherwise hidden structures and relations from text data by grouping similar documents into clusters. The nature of CI activity is an excellent fit for a system able to bring together document clustering with interactive visual analytics, as it involves serendipity in retrieval, browsing, and exploration of text documents. Such a system should be able to direct the analyst's attention to the most relevant information items and allow for efficient exploration of very large document databases.

In this project we propose a neural network-based document clustering system, which is a genuine content-addressable memory system, meaning it clusters and stores text documents so they can be retrieved based on their content. Through the use of Self-Organizing Maps (SOM's), documents are mapped as points on a semantic map, in a topological order that describes the similarity of their contents. This map is then used as a metaphor to interact with the text corpus, creating an interactive visual analytics tool, allowing for its exploration through different, but correlated views and iteratively selecting and examining documents. This means the user can browse through point-and-click, brushing and linking between different views and graphical features or by using prototype examples (e.g. "all the documents similar to this"). An example of this type of environment can be found in our previous project GeoSOM (Henriques et al. 2012), developed by the NOVA IMS team, for clustering and interactive exploration of geo-referenced data. Recent research in text feature extraction, through the use of algorithms such as doc2vec to produce word embeddings, created new opportunities in document clustering. Doc2vec, an adaptation of word2vec, is an unsupervised algorithm that generates vectors for sentence/paragraphs/documents, ordering the documents and semantic information from the text corpus. There are many different word embedding models that, like doc2vec, can convert

more than one word to a numeric vector. These advances in the mapping of words and documents to vectors significantly improved the quality of feature extraction in text mining. The possibility of combining word embeddings and topology-preserving maps (i.e. SOM's) represents a new and exciting research direction. The possibility of coupling improved semantic properties with clustering and visual analytics will allow the creation of effective interactive visual analytics systems, improving access, while supporting serendipity in retrieval, browsing, and exploration of text documents. The interactive visual analytics system proposed here, MapIntel, will represent a significant improvement over the present situation at AICEP and a relevant scientific contribution in the domain of text mining and visual analytics. The ability to cluster and make sense of very large amounts of text documents will not only improve the quality of the strategic planning and decision-making process, but it will also optimize resources by freeing analysts to focus on analysis and adequately briefing policymakers.