Thousands of news articles are published everyday, indicating an extensive media ranging from sports and politics to culture. Some news articles appear isolated, thus without any context to help the reader. But others are well contextualized and seem to initiate or continue a particular story. Due to this large amount of published news articles, readers are constantly struggling to keep themselves updated. Nevertheless, they hardly can avoid losing important news points in the story, which are relevant. “Connecting The Dots Between News” aims to provide a new and richer representation of news stories. With this representation (see image below) the reader can navigate all along the news story and see its evolution within a context.

This work aims to automatically connect news. There are recent studies that focus on connecting the news by topics [1]; and by their subject [2][3][4]. Our approach is composed by three main steps: (i) verify the similarity between news articles; (ii) extract keywords from news and (iii) create the news chains.

Similarity aims to aggregate duplicated or almost duplicated news in groups of news. Such news are those that describe the same event, but presents textual differences, such as: “Nobel da Literatura Gabriel García Márquez hospitalizado no México” - (from Jornal Público - April 03, 2014) and “Gabriel Garcia Márquez hospitalizado no México” - (from TSF - April 03, 2014). Our dataset comprehends 4 million of portuguese news, available online from approximately 50 online media sources (like Jornal Público, Jornal Expresso, Jornal de Noticias and others), and published between 2008 until now (June 2014). Similarity between news has been calculated by means of the Levenshtein\(^1\) edit distance metric. The calculation of the similarity was performed taking into account the time factor to delimitate the news to be compared, and taking three fields into account: title, content and teaser. Based on such similarity ratio, we identify two approaches to identify pairs of similar news articles, a manual one, based on decision trees, and an automatic one, based on Support Vector Classification. The first led to the creation of a decision tree, where the decision values were obtained by experimental observation. The method leading to better results was - Support Vector Classification - with a F1 score of 98.7%.
Keywords can be considered words which can describe what the whole news article spoke in just a few words. We start by extracting keywords from the news content. We considered two types of keywords: simple keywords (consisting of single uni-grams); and compound keywords (n-grams; derived patterns obtained from the POS tag). Keywords were extracted using the TF-IDF (term frequency–inverse document frequency) approach. In addition, entities identified in the news were also extracted, such as names of people. For instance, “escritor” can be considered a simple keyword; “escritor colombiano” a compound keyword and “Gabriel García Márquez” an entity. After performing an error analysis over the extracted keywords, we conclude that the quality of compound keywords is higher than simple keywords, mostly due to the fact that simple keywords are very generic descriptions of a particular news article. The evaluation of keywords was done through a manual verification of whether the words extracted from the news were representative. The percentage of news that were correctly represented are: 73.2% through simple keywords, 76.2% through compound keywords and 80.4% through recognized entities.

The generation of the news chain aims to establish connections between related news groups (that are groups of similar news). Related news groups are groups of news belonging to the same story that appear dispersed along the time. In our approach we decided to use the three mentioned methods plus personalities and the news categories. The personalities were obtained using an external source of knowledge - Verbetes [5] that provides the personalities list mentioned in the news - related to entities and compound keywords extracted. Categories were obtained through tags assigned by journalists to the news with an external source of knowledge that allows to know which category of high-level is associated with each tag. Through the analysis of simple keywords, entities and personalities, words co-occurrence and their respective degree of importance in the group were taken into account. Q-grams² edit distance was used for comparing compound keywords.

For this approach it was considered that only news groups from the same category could establish connections with the other groups from the same category (news groups of Culture are only comparable with news groups of Culture). To establish connections between news groups among other approaches, we used restrictions regarding the number of keywords in common between two groups of news (such as a minimum of three entities in common between news groups) and we also used learning methods. As a consequence of the application of the method, the obtained results lead to the addition of more constraints regarding each news category. Until now, we did a superficial evaluation, my manually evaluating the results. However, we still need to have a more representative evaluation.

An example, of connected news in a story line, obtained by the proposed method, can be observed in Figure 1 where we can see the following set of titles: “Nobel da Literatura Gabriel García Márquez hospitalizado no México” (April 03, 2014) - “Gabriel García Márquez recupera no hospital” (April 04, 2014) and “García Márquez: Cremado em cerimônia privada” (April 18, 2014) - “Juan Manuel Santos pede ‘minuto de silêncio’ pela alma de Gabriel García Márquez” (April 19, 2014).

[4] R Kumar, U Mahadevan and D Sivakumar. A Graph-Theoretic Approach to Extract Storylines from Search Results, 2004