# CREATING AND ANALYSING A SOCIAL NETWORK BUILT FROM CLIPS OF ONLINE NEWS

Álvaro Figueira[1], José Devezas[1], Nuno Cravino[1] and Luis-Francisco Revilla[2]
*arf@dcc.fc.up.pt, jld@dcc.fc.up.pt, nuno.cravino@dcc.fc.up.pt, revilla@ischool.utexas.edu*

*[1]CRACS, INESC TEC & Universidade do Porto*
*Rua do Campo Alegre - 4169-007 Porto, Portugal*

*[2]School of Information*
*1616 Guadalupe Suite 5.536*
*Austin, TX, 78715-1213*
*+1-512-232-2983*

**ABSTRACT**

Current online news media are increasingly depending on the participation of readers in their websites while readers increasingly use more sophisticated technology to access online news. In this context we present the Breadcrumbs system and project that aims to provide news readers with tools to collect online news, to create a personal digital library (PDL) of clips taken from news, and to navigate not only on the own PDL, but also on external PDLs that relate to the first one. In this article we present and describe the system and its paradigm for accessing news. We complement the description with the results from several tests which confirm the validity of our approach for clustering of news and for analysing the gathered data.

**KEYWORDS**

Online news, Web clipping, Social Network, News tagging, Network Graph.

## 1. INTRODUCTION

News media are at a point of historic transition. The conjuncture of digital and social media is merging the roles of readers and providers. Readers are increasingly participating in the news media cycle, where news are written, published, commented, associated with other news, improved, and published again. More than ever, the future of news depends on harnessing the participation of readers in the global process of production and consumption of news.

Many new sites and bloggers try to bind their readers to the news sites using different approaches and means. Usually, the comments are the most participated but, the "like" or the "+1" used in social networks are also increasing their spread among news media. Other less, participated means are the connections to and from blogs, or even the editing of short stories as in Scoop (www.scoop.it). Given this quantity of links, connections and information provided by readers, it would be losing an opportunity if news producers wouldn't take advance of this information in order to better understand reader's needs, interests in order for further development of the stories or even improvements.

In this article we present and describe the "Breadcrumbs" system whose goal is to capitalize on the participation of the general public in the production of news by creating bridges between online news and the "Social Web" while stepping over the traditional techniques of natural language processing and its associated complexity. The project builds on the use of Social Web tools that we created for gathering the opinions of readers for the news that are interesting to them, and then creating a semantically organized model of the readers' opinions. In particular, Breadcrumbs focuses on:

1) Collecting news fragments from the Web;
2) Semi-automatically organizing those fragments;
3) Aggregating the fragments across readers and building the social network of news;

4) Anonymously inferring relationships between readers

5) Inferring relationships between news.

In order to accomplish these tasks, we use various inference and interaction approaches. We combine automatic and user-mediated approaches to yield better results than either approach in isolation – our rationale is: automatic mechanisms can handle extremely large amounts of data, and people can provide insights which are difficult to identify with automatic mechanisms.

The remaining of this article is structured as follows: in section 2 we describe our insight in creating the Breadcrumbs system and try to ground our beliefs in what concerns the viability and opportunity of a social network formed by news clips. In section 3 we detail some of the most important concepts to prepare the reader to understand the state of the art in the area. Section 4 is devoted to the full description of our system, which is then analysed through several tests described in section 5, along with our findings. In section 6 we present our conclusions.

## 2. THE BREADCRUMBS PARADIGM

As evidenced by the success of social bookmarking systems (e.g., delicious.com), people like to keep track of digital information items, storing and collecting them, such that they can be accessed, reviewed, or used later. In Breadcrumbs we extend this by allowing people to keep track of news at a highly fine-grained detail level. Breadcrumbs lets readers select news stories fragments from any news site, blog, etc., collect and store them in their own "Personal Digital Library" (PDL). The fragments can also be annotated with tags and comments. While we feel that currently people are more willing to use tags for better content discovery later on, the comments can provide some sort of metadata that will help the user understand why the fragment is important or why it was actually collected. Therefore, this information is used by the user to create his/her own organization of the clips, or to provide the clips with some kind of mental model for his/her views of the PDL. Concurrently with this, Breadcrumbs also used the tags and comments to learn new characteristics of the collected clips in order to enhance the automatically clustering of the PDL and to infer relations at the system-wide PDL level.

Actually, while each PDL represents the individual perspective of a reader, we believe that by aggregating the PDLs of all readers it is possible to identify previously unavailable patterns and relationships of these perspectives. More specifically, Breadcrumbs organizes the user-selected fragments at the PDL level, and then aggregates PDLs at the system-wide level using text mining and social filtering techniques.

In order to organize each PDL for each user, the system applies an automatic mechanism that clusters the news fragments based on their content and on the semantic proximity between them. As for the PDL aggregation, we focus on text mining, community detection and, social classification methods (Gibson, 1998) that potentially identify implicit links or relationships between fragments based not only on text similarity, but also on the tags and on the comments assigned by the users.

As a result, we expect to create a social network based on these implicit links. While the idea of social network is usually associated with services such as Facebook or Google+, where people connect with other people, a social network doesn't necessarily represent direct person-to-person connections, but rather some kind of socially-induced relationships (e.g. social bookmarking, where relationships are established between web sites based on user-assigned tags). Breadcrumbs relationships exist between news clips, not people, but they are nevertheless influenced by human behavior, thus following the basic characteristics of traditional social networks.

The collected text fragments are mined to discover entities by comparing them with DBpedia's entries which allow us to, independently from the language, be able to structure the information. In particular, we are able to detect people, places and time periods. Then, we build a network of co-occurences of entitites which ultimately forms a graph on a multi-level dimension.

Summing up, with the features and tools, we believe this built network, and its correspondent graph, will allow journalists and news agencies to: learn which stories and wordings resonate with the readers; identify the hot and most "trendy" topics, and identify previously undetectable connections between apparently disconnected information sources; finally, the system can identify user communities; and provide users with further reading suggestions.

We also believe that Breadcrumbs has potential to create a new paradigm in information sharing and for navigation in online news. Readers are able to organize their local PDL, but with Breadcrumbs the information stored (anonymously) is freely available to all users of the system. Navigation on the fragment-graph is therefore based on automatic inference laws provided by the relations between fragments, and by the local organization of the clips, which in turn is influenced by the social classification elements that were involved during the creation of the clusters of web clips (the assigned tags and comments).

## 3. TECHNICAL BACKGROUND

As described in the previous section, Breadcrumbs basically focuses on: 1) organizing news fragments collected from the Web, by readers; 2) inferring relationships between readers, and inferring relationships between news.

The first focus is mainly a classification problem that lies in the area of text mining, and information retrieval. However, in Breadcrumbs we also propose a solution to solve the problem of integrating social classification elements with standard clustering algorithms. In this system we use the news fragment as the core information element to be considered. Without loss of generality, an online news fragment can be considered a "document" which is the term often used in research in the area, and that we will use in this section hereafter.

Information retrieval strategies usually assign a measure of similarity between a query and a document. These strategies are based on the common notion that the more often terms are found in both the document and the query, the more relevant the document is deemed to be. Some of these strategies also employ "counter measures" to alleviate problems that occur due to the ambiguities inherent in language and also to different semantic definitions for terms. In a Vector Space Model (VSM), one would compute a measure of similarity by defining a vector that represents each document, and a vector representing the query, as it was initially presented (Salton et al., 1975). This model is based on the idea that the meaning of a document is conveyed by the words it uses. However, this strategy assumes that every word has the same importance for classification. Robertson and Spark Jones, (1976) showed the importance of having a collection of terms with weights and their research led to a formula for weighting terms involving the term frequency (TF) and the inverse document frequency (IDF) frequently known as TF-IDF, which is the current general basis for weighting words. This formula has been improved several times (Salton and Buckley, 1988) and one of the latest versions is due to (Singhal, 1997) in which the relevance of each document is also a function of its size. Latent Semantic Indexing (LSI) (Deerwester et al., 1990) is an improvement for information retrieval for situations when there is a direct map into keywords (as it is the case of VSM). Since the same concept can be described using many different keywords, this type of matching is prone to failure. A recent technique to deal with the problem uses Singular Value Decomposition (SVD) to filter out the "noise" found in a document such that two documents that have the same semantics (whether or not having the same terms) will be located close to one another in a multi-dimensional space. Dumais (1994) showed that LSI performs slightly better than conventional VSM. However, the run-time performance of LSI is a serious concern. The SVD is itself computationally expensive (for *n* documents, and a matrix of rank *k*, an $O(n^2 \times k^3)$ algorithm is available).

On the other hand, social classification can be seen as a "Relevance Feedback" (RF) problem, which tries to identify the documents that are deemed to be relevant either by manual intervention or by assumption that the top documents are relevant. The adaptation of RF to VSM was done several decades ago (Rocchio, 1971), but only 21 years later (Harman, 1992) it was investigated the effect of using multiple iterations of relevance feedback. Three years later, Spink (Spink, 1995) suggested a method in which the user intervenes in picking and choosing which terms to be relevant. This last approach is clearly the most similar to the one taken in project Breadcrumbs.

Document clustering attempts to group documents by content to reduce the search space required to respond to a query. Different clustering algorithms have been proposed, but in all of them the efficiency factor is present because of their computational complexity from a temporal and space point of view. One of the most popular ones is the K-Means (Willet, 1990) a partitioning algorithm that iteratively moves *k* centroids until a termination condition is met (usually until the centroids do not move anymore). The Buckshot Clustering Algorithm (BCA) was designed to run in $O(k \times n)$ time, where *k* is the number of clusters and *n* is the number of documents. Therefore, it is an interesting improvement over alternatives that require a document-document similarity match, that run in O(n$^2$). Details of the BCA and its analysis are described in (Cutting et al., 1992). Some studies have found that hierarchical algorithms, particularly those that use group-average cluster merging schemes, produce better clusters. However, more recent work indicates that this may not be true across all metrics and that some combination of hierarchical and iterative algorithms yields improved effectiveness (Steinbach et al., 2000; Zhao and Karypis, 2002). Yet, as these studies used very small document collections, it is difficult to conclude which clustering method is definitively superior. In Breadcrumbs we propose an algorithm capable of similar (asymptotic) performance to k-means.

When trying to establish links between news and readers it is also important to have a sense of “relevance” or of “authoritative”. The Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg (Kleinberg, 1997). It determines two values for a page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages. Hubs and Authorities is a scheme used for ranking web pages based on the idea that certain web pages, known as hubs, serve as large directories that are not authoritative in the information that they held, but are used as compilations of a broad catalog of information that lead users directly to other authoritative pages. The scheme therefore assigns two scores for each page: a hub score and an authority score. This model was adopted, and adapted, in Breadcrumbs in order to determine the Hubs in the network as leading trails for the discovery of news usage and duplication.

Another important social network analysis area that contributed to the Breadcrumbs social classification problem was the identification of community structure. The term “community” was coined from sociology and can mathematically be translated to “dense subgraph in a social graph”, that is, a set of people (nodes) that have more connections between themselves than they do with the remaining network. Community detection algorithms are currently being actively researched, as several scalability and community quality problems still exist. Suprise maximization (Aldecoa and Marín, 2011) is one of the most promising algorithms in this area. On the other hand, Tang et al (2011) have reused traditional community detection algorithms, including Latent Space Models, Block Model Approximation, Spectral Clustering and Modularity Maximization to propose a unified view that allows the application of these methodologies to multidimensional networks (several types of edges with different relationship information), integrating dimensions during one of four possible stages. Although the idea of community detection is usually associated with a network of people, it can be applied to other types of socially-induced networks, such as folksonomies, where tags are assigned by users to classify documents. In the Breadcrumbs system, we share this same perspective, applying community detection to networks of “social documents” as opposed to networks of people.

## 4. DESCRIPTION OF THE SYSTEM

Breadcrumbs is an integrated system based on four components: 1) a Clipper to assist the user while collecting, tagging and commenting textual fragments from online news, and also to store them on the PDL; 2) a Proxy to provide the Clipper functionality in every browser in a transparent way to the user; 3) a Personal Digital Library (PDL) where the collected clips are stored and automatically organized. The PDL can be accessed later on and re-organized, if wanted; 4) a Classification and Clustering Engine (CCE) to organize the clips automatically in clusters, based on semantic proximity of content and tags/comments, and to infer relations between the clips; 4) a Social Graph that connects the clusters in each PDL according to the discovered relations that exist between clips. This Social Graph can be further explored as is described in section 4.5.

In the following sections we describe each of these components. We hide some implementation details about authentication procedures, session maintenance and access to repository and database, focusing primarily on the observable functional behavior.

### 4.1 The Clipper

A key aspect component of the system was to allow users to select textual fragments from online news sites or blogs, and collect them in their PDL for later access and manage, in the least demanding way.

For this, one crucial issue was to make sure any modern browser could be used for this task. We created a light-weight interaction system that allows users to collect fragments without the need of any plug-in. From the point of view of the user, the collection mechanism is comprised of a set of very small and intuitive steps. The clipper itself works from the personal area of Breadcrumbs system. Once there, the user finds an entry box (in the Google style) to enter the desired URL or web address of the site he/she want to visit. Therefore, the actual steps for collecting some text fragment from the web are:

1. enter a link in a text box and press enter to navigate to the intended web page (another possibility would be to accept one of the automatic personalized suggestions of links/pages to navigate to, or even t just select one link form a list of the most used links);
2. select the text to be clipped by using the standard procedure to select text in any modern browser using the cursor;
3. press the ‘plus’ button in the Breadcrumbs dashboard that appears in the top left of the window. Notice that this dashboard is created with code injected in the web page by our dedicated proxy.

In Figure  we compare two situations: on the left we present a news web page, as any reader would see it, but with the clipper dashboard on the top left corner. The presence of the dashboard means that the system is running on the background and is able to assist the user in the task of clipping text fragments. On the right side, we see the same page, but now, having seven clips collected and listed in the dashboard drop list. This happens because the system detects that the user has already seven clips collected from the same URL.

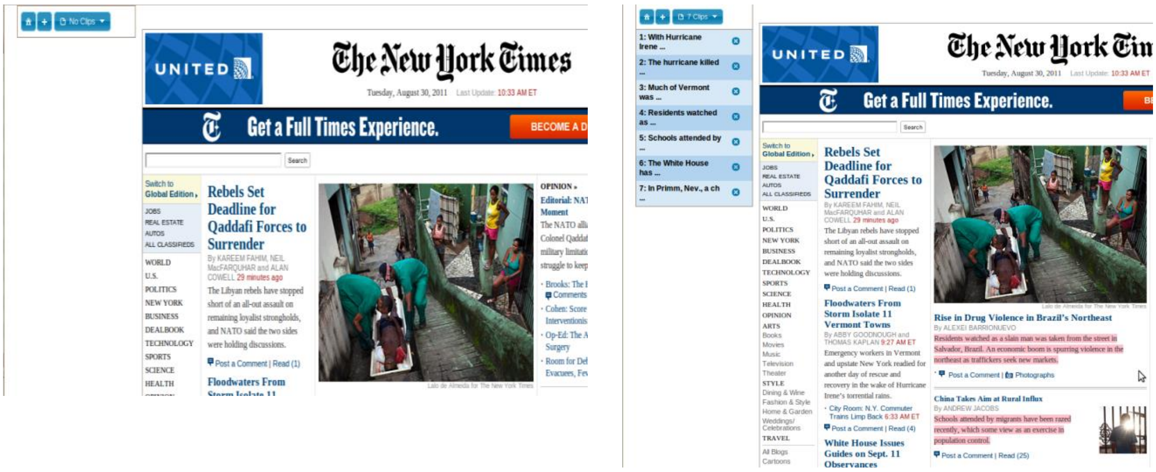

No Clips
The New York Times
Get a Full Times Experience.
Rebels Set Deadline for Qaddafi Forces to Surrender
Floodwaters From Storm Isolate 11 Vermont Towns
7 Clips
1: With Hurricane Irene ...
2: The hurricane killed ...
3: Much of Vermont was ...
4: Residents watched as ...
5: Schools attended by ...
6: The White House has ...
7: In Primm, Nev., a ch ...
Rise in Drug Violence in Brazil's Northeast
China Takes Aim at Rural Influx
White House Issues Guides on Sept. 11 Observances

**Figure 1. Collecting web clips.**

The collecting tool is based on an adaptation of the target web page through a dynamic injection of JavaScript as we describe in the next section.

## 4.2 THE PROXY

Our approach in building the system was to have the clipper running on the client-side, mainly for efficiency reasons. Anytime a user visits one web page (not protected by password) he/she gets the intended web page plus the clipper dashboard on the top left, as we have seen above. To achieve this functionality we use an in-house developed proxy that passes transformed web pages to the client system, which is now equipped with a dashboard to perform the clipping operation. Hence, every link in the visited web-page is also transformed into a special request to the proxy in order to process that URL.

The interactions between user requests and the system components are illustrated in the following UML diagram presented in Figure . Although the global sequence is transparent to the user, the system still needs to ensure that every clip has the right owner (we use HTTP sessions), and that the clip is stored in the right PDL, which in turn is stored in the Breadcrumbs system repository.

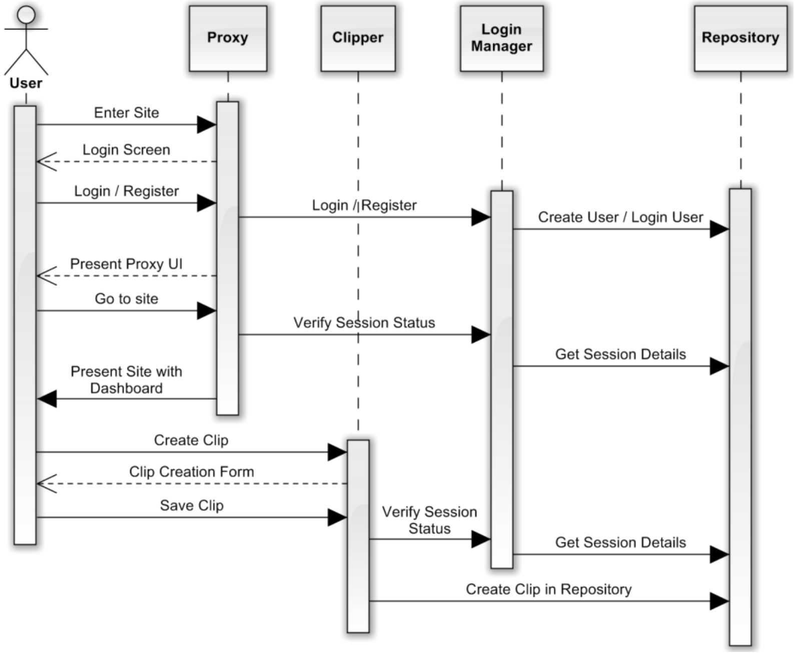
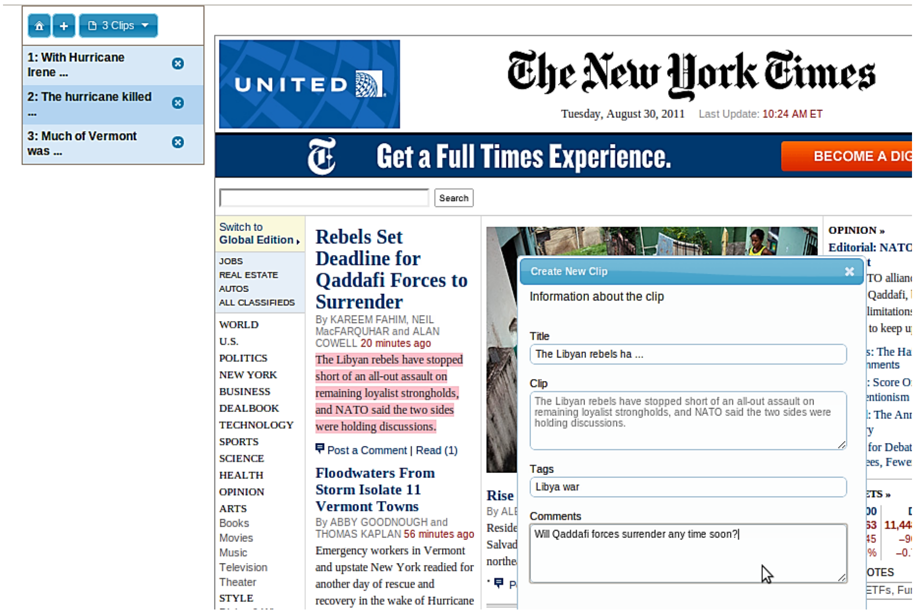


**Figure 2. Internal clip collecting sequence.**

Therefore, the proxy is totally transparent to the user as he/she can browse the web, navigating from one link to another without ever noticing that all web addresses are being translated and that every visited web page is being injected with code to produce the dashboard on the client side.

During the clipping operation the user may tag and comment the current clip. This situation is illustrated in Figure  where after selecting the text to clip and pressing the ‘plus’ button on the dashboard, the user is presented with a window. This window offers to the user the possibility of adding tags to the selected text; and also to add a comment, before the confirmation of the clipping operation.



Rebels Set Deadline for Qaddafi Forces to Surrender

By KAREEM FAHIM, NEIL MacFARQUHAR and ALAN COWELL 20 minutes ago

The Libyan rebels have stopped short of an all-out assault on remaining loyalist strongholds, and NATO said the two sides were holding discussions.

Post a Comment | Read (1)

Floodwaters From Storm Isolate 11 Vermont Towns

By ABBY GOODNOUGH and THOMAS KAPLAN 56 minutes ago

Emergency workers in Vermont and upstate New York readied for another day of rescue and recovery in the wake of Hurricane

Create New Clip ×

Information about the clip

Title: The Libyan rebels ha ...

Clip: The Libyan rebels have stopped short of an all-out assault on remaining loyalist strongholds, and NATO said the two sides were holding discussions.

Tags: Libya war

Comments: Will Qaddafi forces surrender any time soon?|

**Figure 3. Tagging and commenting a clip.**

While tagging and commenting clips is an operation that can be done during collecting time, it also can be done at any time later. For such, the user only needs to access her/his PDL; pick the clip from the workspace and fill/change the tags or the comments fields. The Breadcrumbs tags can be freshly created and are not restricted to any closed vocabulary.

## 4.3 The Personal Digital Library

Traditionally, people tend to use some form of classification to organize their personal data. We took the view that the same would occur if we give readers the tools to collect and classify the fragments in the Personal Digital Library (PDL). We extended that classification to make it "semi-automatic" in the sense that, the classification mechanism uses the fully automatic inference system based on clip content, plus the data entered by the user in the form of tags and comments. Conceptually, this situation is illustrated in Figure where a user is capable of adding tags/comments to its personal digital library. This figure also depicts the feature that whenever a new clip is inserted into the PDL it is automatically organized in an appropriate cluster of other clips.

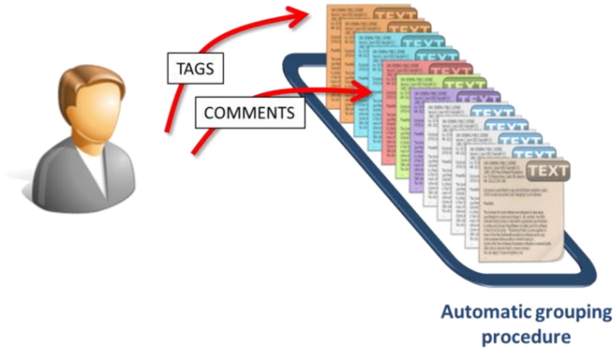


**Figure 4. A user tagging and commenting his automatically organized web clips.**

The PDL is an area where each user can see and manipulate the collected clips. It I comprised of a toolbar and a workspace, as presented in Figure 5. The collected clips are depicted in the workspace using rectangles with a blue bar on top. Clips may stand on piles or tiled. The user can also create manually clusters (groups) for a set of clips, or may create labels to associate with clips. This strategy is strongly based on human-usual behavior. When we pile a set of documents we are implicitly grouping then because of some common characteristic. Breadcrumbs is capable of detecting piles of clips and create a manual cluster from that information. The same would occur if the documents are put in some kind of a tiled organization.

On the other hand, it is possible to ask the system to automatically organize the workspace (i.e., all clips there were not manually grouped with others). We call this feature asking for "recommendation of organization". We have also the possibility to trigger a parsing of the workspace whenever the user wants, to detect the new piles and tiles of clips.

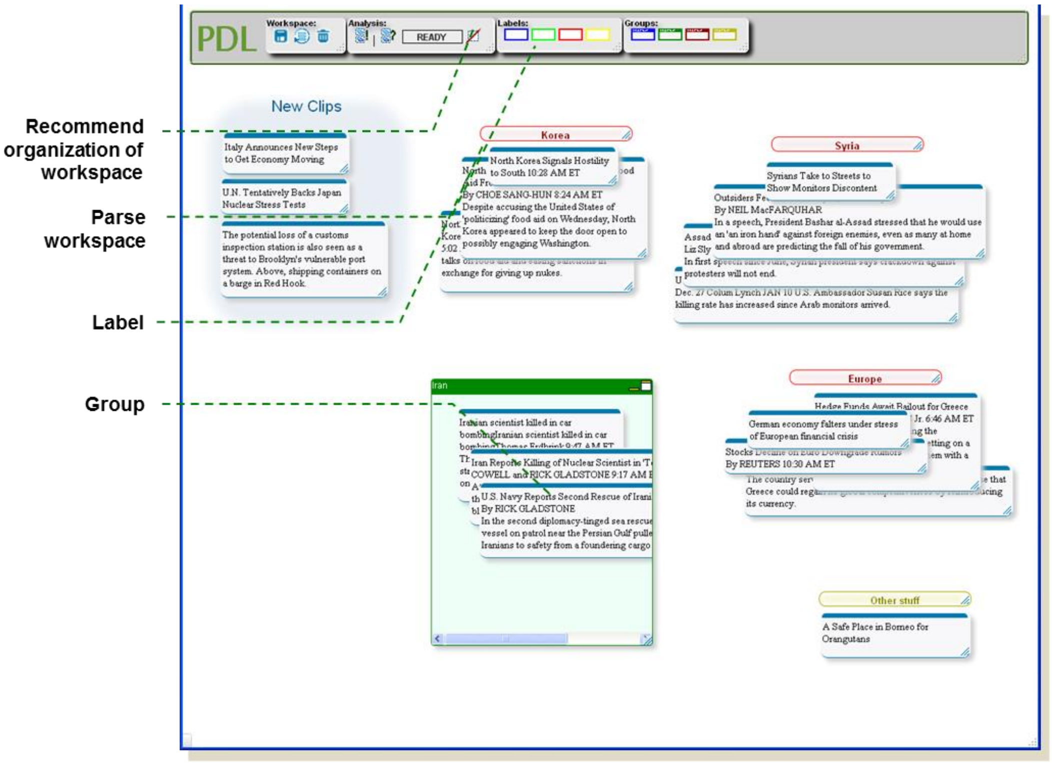

Recommend organization of workspace
Parse workspace
Label
Group
PDL
Workspace:
Analysis:
READY
Labels:
Groups:
New Clips
Italy Announces New Steps to Get Economy Moving
U.N. Tentatively Backs Japan Nuclear Stress Tests
The potential loss of a customs inspection station is also seen as a threat to Brooklyn's vulnerable port system. Above, shipping containers on a barge in Red Hook.
Korea
North Korea Signals Hostility to South 10:28 AM ET
By CHOE SANG-HUN 8:24 AM ET
Despite accusing the United States of 'politicizing' food aid on Wednesday, North Korea appeared to keep the door open to possibly engaging Washington.
exchange for giving up nukes.
Syria
Syrians Take to Streets to Show Monitors Discontent
By NEIL MacFARQUHAR
In a speech, President Bashar al-Assad stressed that he would use an 'an iron hand' against foreign enemies, even as many at home and abroad are predicting the fall of his government.
protesters will not end.
Dec. 27 Colum Lynch JAN 10 U.S. Ambassador Susan Rice says the killing rate has increased since Arab monitors arrived.
Iran
Iranian scientist killed in car bombingIranian scientist killed in car
Iran Reports Killing of Nuclear Scientist in
COWELL and RICK GLADSTONE 9:17 AM
U.S. Navy Reports Second Rescue of Irani
By RICK GLADSTONE
In the second diplomacy-tinged sea rescue vessel on patrol near the Persian Gulf pulle Iranians to safety from a foundering cargo
Europe
Hedge Funds Await Bailout for Greece
German economy falters under stress of European financial crisis
Stocks Decline on Euro Downgrade Rumors
By REUTERS 10:30 AM ET
its currency.
Other stuff
A Safe Place in Borneo for Orangutans

**Figure 5. A screenshot of the user's PDL.**

In order to facilitate the process of integrating new clips into the existing organization, PDL was augmented with adaptive mechanism that recommends where to move individual clips. The user asks PDL to suggest a location by clicking the button at the top of Figure 5. At this point, PDL invokes the spatial parser, which identifies 5 clusters (4 piles and 1 independent clip). The recommended location for the clip is determined by computing the similarity between the clip's text and metadata and the aggregated text and metadata of the spatially inferred groups in the workspace. In this case, the similarity between the texts is very low, but the clip's tag is very similar to the tag cloud for the "Europe" pile. PDL conveys its recommendation by sliding the clip to the Europe pile. In order to accept the recommendation, the user must click on the clip. In this case, the user does not accept the recommendation, so the clip bounces back moving quickly back into its original position.

After moving the new clips into various groups, the user starts considering if these new additions to her collections might have changed the overall nature of the collection, making it necessary to reorganize the whole workspace. By collaborating with the CCE, the PDL can recommend alternative organizations for the whole workspace. The user request PDL to recommend an overall organization of the workspace by clicking the button at the top of Figure 5.

At this point, PDL asks CCE for the best categorization of the clips. CCE returns a one level set of clusters. Based on this, PDL makes a suggestion by animating all clips, such that they move into a new location.

## 4.4 The Classification and Clustering Engine

The Breadcrumbs system automatically organizes fragments in each PDL according to their semantic proximity which in turn is computed from: a) the fragment's content; b) the social elements associated with each fragment. This organization is performed by the Classification and Clustering Engine (CCE) module. The classification of these fragments (i.e. documents) is a process of putting each fragment in a cluster with includes all the other fragments that are semantically related to the first one. We expect that this classification will simplify a search for past collected news fragments, and that will provide her/him with new leads for further reading.

The algorithm to integrate the two classification schemes is based on the popular k-means (Willet, 1990), adapted to include tags/comments. Unlike many other approaches (Ares et al, 2011; Ji et al, 2006; Tang, 2010), we take the view that tags may stand outside of the clustering process of the documents and form on their own a, eventually overlapping, community structure. The discovery of this community structure of tags is the first step towards the integration of this social part in order to enhance text clustering. In our approach (Cravino et al, 2012) we define a distance metric based on a weighted cosine similarity, which combines the textual features with the community structure of a network of tags in order to improve the clustering of documents in a socially biased way.

We use the Speaker-listener Label Propagation Algorithm (SLPA) proposed by Xie et al (2011) to identify the cover (the overlapping community structure) of our network of tags. The SLPA algorithm is mainly used due to its near linear complexity in sparse graphs. This algorithm works by propagating labels throughout each node of the network that are repeatedly stored in memory for every node. It has two parameters, a threshold of probability, used in the post-processing phase, and the number of iterations. After an initialization step, SLPA starts by taking each node in a role as listener and receives one random label from each of its neighbors (speakers) which stores in a temporary list. The listener then chooses a label from this list and adds it to its own memory according to a function based on the label occurrence count. The previous process is iterated a number of times according to the parameter, and the node memories are then post-processed. The post-processing step consists of the computation of the occurrence probability for all labels and the removal from node memory of all labels with an occurrence probability below the threshold. The sets of nodes that share a certain label in its memory are constructed yielding the communities of related tags.

We introduced modifications to this algorithm to make it work with a weighted network using a modified labels list to store the sum of weights connecting to the speakers from where the label came from. The listener rule was also modified to return the label with the maximum value for the product of the sum of its weights with its occurrence count.

The CCE is then enhanced by giving each user the possibility to choose the amount of importance given to the tags during classification. We present to the user a ‘Social Sliding’ bar (SS), which ranges between ‘no importance’ of tags/comments during clustering, through ‘only use tags/comments’. Therefore, this parameter allows each user to define a different behavior of the CCE, according to the position of the SS in its scale. The SS parameter influences is integrated in the classification process when computing the tf-idf. We construct the word vector for each clip according to the tf-idf score, and the tag vector for each clip using the following tag weighting function (Cravino et al, 2012):

$$w(t,d) = (1-SS)\times tfidf(t,d) + SS \times \frac{1}{|C_t|}\sum_{tr \in C_t} tfidf(tr,d)$$

Where *w(t,d)* is the computed weight of tag *t* in document *d*, and $C_t$ is the union of all overlapping communities of tags related with tag *t*.

The algorithm equipped with this formula allows a dynamical perspective from the user side in the sense that the user may choose the degree of importance to give to tags when forming the clusters.

### 4.4.1 Integration of Tags into the Clustering Process

We manually annotate the news clips collection, classifying each clip into one of the following six classes: Libya, US Tax, World Debt Crisis, Italy Downgrading, Greece, and Other. We use this clustering partition as our “ground truth”, to which we compare the partitions resulting from the text clustering and from the combination of the text clustering with the tag clustering. In Table 2 we present the confusion matrix analysis for each of the methods, where “class” refers to our manual annotation of the clips and “cluster” refers to the partitions identified by the tested methods.

The true positive rate (TPR) for the text-based method is 32.15% and the false positive rate (FPR) is 26.32%. Even though the TPR for the combined text and tags method takes a lower value of 29.70%, the FPR also decreases to 23.55%, which means that the text-based method achieves a higher number of correctly classified documents, but also a higher number of incorrectly classified documents. Since these metrics do not provide the grounds for a conclusion, we use the Rand index (Rand, 1971) to measure the similarity of the resulting partitions with the ground truth, i.e. the percentage of correct decisions, and the F-score to calculate the accuracy of the two methods, first using $\beta = 1$ and then using $\beta = 0.5$ and $\beta = 2$ to penalize the false negatives less and more strongly, respectively, than the false positives.

Table 1 depicts the evaluation of the identified partitions using a null weight (Text), as well as a 50% weight (Text+Tags) for the social aspect. That is, for the Text clustering, we set the social slider to zero (SS = 0), while for the Text+Tags clustering we set the social slider to 0.5 (SS = 0.5), in our weighted cosine similarity proximity measure. As we can see in Table 3b, we obtain a higher Rand index when using the social structure of the network of tags in the clustering process. On the other hand, by looking at the F-score for either method in Table 1a, we verify that using the community structure of the co-occurrence of tags in news clips slightly decreases the accuracy of the clustering method, except when given a higher weight to the precision ($\beta < 1$), being consistent with the changes in the values of Precision and Recall elicited by the choice of clustering method shown in Table 1c. Since the F-score values for the two clustering methods are very close together and the Rand index isn't by itself conclusive, we further investigate by calculating the adjusted Rand index according to Hubert & Arabie (1985) and Morey & Agresti (1984). These adjusted for chance metrics are depicted in Table 3b. The resulting values are in agreement with the previously calculated Rand index, indicating that the higher Rand index for the Text+Tags clustering represented in fact a significant result.

**Table 1a**

| **Clustering** | **F0.5** | **F1** | **F2** |
|---|---|---|---|
| Text | 0.244 | **0.268** | **0.298** |
| Text+Tags | **0.246** | 0.263 | 0.282 |

**Table 1b**

| **Clustering** | **Rand Index** | **HA ARI** | **MA ARI** |
|---|---|---|---|
| Text | 0.654 | 0.050 | 0.075 |
| Text+Tags | **0.672** | **0.056** | **0.082** |

**Table 1c**

| **Clustering** | **Precision** | **Recall** |
|---|---|---|
| Text | 0.230 | **0.321** |
| Text+Tags | **0.236** | 0.296 |

The socially biased document clustering method that we've introduced here was able to produce an improved clustering partition, by taking advantage of the social features in our documents. Identifying the overlapping community structure of the network of tags associated with the Breadcrumbs folksonomy seems to improve regular text clustering, resulting in a better grouping division of our news clips collection. We hypothesize that, as the network of tags grows and its community structure becomes stronger, groups of tags will become more cohesive and continuously result in improved socially biased clusters.

### 4.4.2 Assigning Topics to the Clusters

After the clusters are identified, we take advantage of Latent Dirichlet Allocation (LDA) (Steyvers, 2007) to train a topic model that can be used to the identify the subject or title of each cluster.

LDA is a probabilistic generative model, that expresses the idea that documents can be created from a combination of different topics, where each topic has a different weight (or probability) in the document. LDA is able to capture the semantic of documents and to create different topics for homonym words (e.g. it is able to differentiate between "bank/economy" and "river bank", capturing the different meanings of the word "bank" depending on the context).

The training set used to generate the topic model consists of the Breadcrumbs corpus of news clips, as well as an additional corpus comprising full news articles from the Reuters news agency.

The Reuters-21578, Distribution 1.0 test collection is available from David D. Lewis' professional home page, currently: http://www.daviddlewis.com/resources/testcollections/reuters21578/. We generated a topic model with 400 topics and used 2000 iterations to optimize the model.

Using Latent Dirichlet Allocation, we were able to provide a service within the Breadcrumbs system capable of automatically assigning an appropriate title to each group of news clips, displayed in a user's PDL.

## 4.5 The Social Graph

The Breadcrumbs system aggregates and organizes news fragments at two levels: at the PDL level and at the Social Graph (SG) level. PDL classifies fragments in clusters. The SG encodes the aggregation and inference results much in the same line as described in Figueira et al (2007) for “iGraphs”. Therefore, our interface is designed to allow readers to navigate and interact with the PDL and, to provide a means to navigate and interact with external clusters (external to the PDL) of clips which would ultimately present new perspectives of a story and different topic associations. In this last case we are able to learn which stories and wordings resonate with other readers.

### 4.5.1 The Nodes and Edges of the Social Graph

By aggregating all PDLs in the system we can establish relations between fragments. These connections between the collected fragments are based on ‘strong relations’ (we consider them “equivalence of fragments” when most of the text is shared between the fragments), ‘weak relations’ (fragments collected from the same web page but that not overlap), ‘transitive-strong relations’ (equivalence of two fragments through a third fragment that encloses both) and ‘transitive-weak relation’ (which includes fragments collected from the same web domain, or news publisher). In Figure we illustrate the strong and weak relations. Note also that in this figure all the three clips share a transitive-weak relation because they all are collected from the same web page.

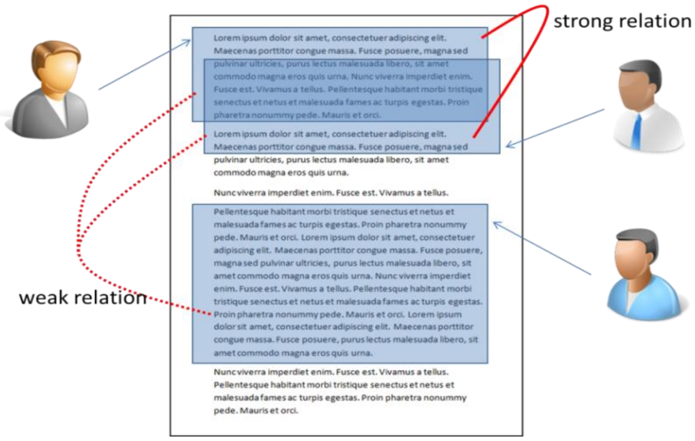


**Figure 6. Relations between web clips.**

We may now take the view that the clips and the connections that exist between them form a network and associate this representation to a graph, being the clips the nodes and the connections the edges of the graph. This constructed network can then be navigated using a system capable of depicting the fragment-graph and allowing the access to the ‘remote’ fragments through the network links. These remote fragments are therefore, the clips that lie in the same remote cluster, which has (at least) one clip that is related to some local one.

The system also offers to the user a vision of news fragments in several “domains”: own PDL, remote PDL (PDL managed by another user); in a local cluster, in an external cluster; in the same web-page; or in the same domain.

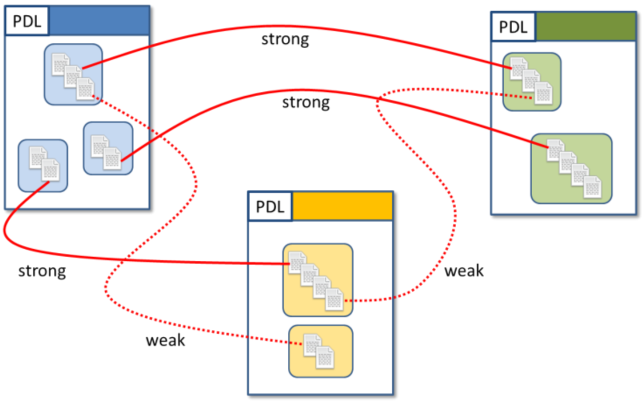


**Figure 7. Connections between clips – the Social Graph.**

### 4.5.2 Using the Social Graph

From the created graph we are able to identify important properties in a network of news fragments which is a similar task to the one described by Gibson (1998). We can also understand how different PDLs may connect through implicit relations that can be established between news fragments, and also understand the importance of ‘hubs’ of information in the transmission/propagation of information in the network.

We included mechanisms to analyze the characteristics of the network in respect to metrics common in Social Network Analysis (Scott, 2000). These include the “centralization degree” and the “centralization index” to find “hubs” of information in the network; the search for “cliques”, especially in the scope of PDLs, would also find a network of common interests among the owners of each PDL. The network “density” is also an important parameter if computed together with “bridges” (edges which avoid the separation of the network in two parts) to identify sub-communities in the network.

We provide a set of tools that explore these metrics and the co-referenciation of named entities by creating a sub-graph that establishes links between fragments that refer to the same named entity. Currently we find the entities by comparing the fragment with DBpedia. This allows us to form a multi-dimensional graph based on people, places and time periods (Devezas et al, 2012).

## 5. ANALYSIS FROM THE SOCIAL GRAPH

We believe that community detection methodologies can similarly be applied to our clips data set, for the analysis of a more complete named entity network, where people, places and dates can all be connected if they are co-referenced in a web clip. Based on this network, we can then find communities, clusters induced by the clipping behavior of people. This feature leverages the capability to provide insights from the context of our corpus, as an attempt to answer the questions “What?” and “Why?” and to emphasize the highly related and densely connected groups of entities.

### 5.1 Clip-centric network versus Entity-centric network

The Breadcrumbs system uses a clip-centric network, where relationships between clips are established by the co-reference of an entity in a pair of distinct clips, as opposed to an entity-centric network, where relationships are established by the co-reference of a pair of distinct entities in a single clip. The clip-centric model has a clear advantage over the other as it enables the direct mapping of results into clips instead of entities, but given the difference in paradigm it is still uncertain whether or not it produces similar groups of information.

In our model, we established a connection between a pair of entities, whenever they were mentioned together (co-referenced) in a clip. Since the entities had been previously resolved to their corresponding URI, we could also say that we are establishing a language-independent context.

We apply this idea to our test set, a collection of 259 news clips, gathered independently by 5 different people, across a period of 24 hours, from five news sources — Washington Post, Times, Telegraph, Guardian and Daily Mail — and covering five main topics — Libya, US Tax, World Debt Crisis, Italy Downgrading and Greece. We limit the ontology-based named entity recognition process to Place subclasses — Country, Continent, Island, Natural Place and Historic Place — and Person subclasses — Politician, Office Holder, Athlete, Cleric, Scientist, Model, Criminal and Judge.

This is a clear indication of the absence of traditional natural language processing methodologies, which usually focus on the grammatical analysis in order to identify the phrase structure, and from there to the identification of the entities. In our approach (Devezas et al, 2012) we have abstained from following this path, as these methodologies are usually language-dependent.

In our view, a news clip will be pertinent to its creator and will possibly contain some of the most relevant information of the news story that he/she collected. However, it's the connection of all this information that will impose meaning and establish the context of a group of news fragments. These groups act as contextual supernodes aggregating smaller nodes with a common topic.

Next, we discuss the creation and analysis of two clip/entity networks based on the co-reference of people and places (the entities) in the news clips of the Breadcrumbs system. We introduce the tools that we used to process our data, describing their general purpose as well as the data flow process between each application. We started with a tabular data set containing the columns *clip* and *entity*, that describe which entity is cited in which clip (e.g. Clip 1 / Barack Obama, Clip 1 / United States, Clip 2 / Barack Obama, etc.). In order to transform this data into a network, we took advantage of the R Project (R Development Core Team, 2011), which is a free software environment for statistical computing and graphics, with several available packages for diverse mathematical and analytical operations. Specifically, we were interested in creating a graph structure, which the igraph package (Csárdi and Nepusz, 2006) enabled us to do. We pre-processed the data from the Breadcrumbs system using the R language and the igraph package, transforming the clip–entity dictionary into two separate networks, one for each network model, that we exported to GraphML (Brandes et al, 2002). We then analysed the generated networks using the Gephi system (Bastian et al, 2009). Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. It enabled us to do an exploratory visual analysis of the networks, computing the eigenvector centrality for every node in the graph, and identifying their community structure using the modularity-based methodology by Blondel et al (2008), implemented in this system.

The resulting network contains 74 nodes and 231 edges, having a density of 8.55% and a diameter of 14. By analyzing the largest component of the graph, we were able to identify three large communities, which are further described in Table 2.

We rank nodes by eigenvector centrality, retrieving the top 5 nodes for each community, to help with topic identification and the validation of the cluster as a language-independent contextual supernode. For the clip-centric network model, we manually assigned keywords that describe the content in each clip. Notice the fact that the same eigenvector centrality is easily explained by the existence of similar connections induced by the same named entity set. We do not assign any keywords to the top five nodes of the entity-centric network, since we use instead the entity label and our personal knowledge about the current world affairs to infer the topic of each community.

We compared the two models (clip-centric and entity-centric) and found them both to be adequate to describe this relational information, given they both present the common characteristics of real networks, having an inherent community structure that enables the identification of what we called language-independent contextual supernodes.

However, the clip-centric model has the advantage of directly mapping the contextual communities into groups of news clips, which then allows for an in-depth analysis of the groups. On the other hand, the entity-centric model proved to be more simplistic, in the sense that it is more reduced and can easily be used to visually illustrate the context of a news corpus, be it the whole news clip collection or the news clips in the personal digital library of a user.

**Table 2. Eigenvectors centrality**

| Community ID | Eigenvector Centrality | Entity Label |
|---|---|---|
| 5 | 1.000000 | Greece |
| 5 | 0.907215 | Italy |
| 5 | 0.901497 | Europe |
| 5 | 0.664476 | Spain |
| 5 | 0.663872 | France |
| | | |
| 7 | 0.182921 | Barack Obama |
| 7 | 0.144418 | The President |
| 7 | 0.131588 | United States |
| 7 | 0.129832 | John Boehner |
| 7 | 0.101335 | Muammar Gaddafi |
| | | |
| 9 | 0.857089 | Libya |
| 9 | 0.701464 | Africa |
| 9 | 0.699071 | Russia |
| 9 | 0.678829 | India |
| 9 | 0.678829 | Jordan |

## 6. CONCLUSIONS AND FUTURE WORK

The problem Breadcrumbs is facing is rather relevant and interesting. On one hand, we need to provide readers with new reading experiences that news reading systems still don't have, on the other hand the scientific challenges are considerable: current classification algorithms are already capable of organizing textual information according to its semantic nature, using text mining techniques. Although machines can process enormous amount of information at a very fast pace, they do not have the insight to infer relations from data if those relations are based on information not present in the data itself.

Currently the Breadcrumbs system has: a) a news collecting tool that stores news fragments in a Personal Digital Library; b) the capacity to semantically organize the PDL content according to automatic and social classification; c) an interface capable of providing readers with an easy way of viewing the PDL content, to classify clips with tags and comments; to search and retrieve fragments of news and also to view the original web page from where the clips were collected. This tool provides news readers with a PDL which may be considered their perspective on the news, and the corresponding organization based on that perspective. The automatic discovery of links between clips lying in different PDLs provides the reader with the capacity to navigate to an external cluster to refer to another perspective, which in turn may have other clips with another set of external links, thus providing a hyper-text navigation of clips that can be browsed by the reader.

In the near future we expect to have a graph with numerous PDLs and clips, and hence the capacity to extract important information from this network graph using a social network analysis perspective, and taking advantage of the collaborative nature that the social aspect provides to document collection and organization. From this panorama we will be able to assess and evaluate a set of other relevant topics: the capacity of the system to inform journalists of important elements used by readers when classifying fragments; the capacity to inform journalists of links between readers and of communities of readers based on their interests; the capacity of the system to inform about value of classification elements; capacity to find a trail of particular fragments in different news sources; the overall satisfaction in using the system. We will also extrapolate our findings trying to infer the importance of the developed system for both readers and journalists.

## 1. ACKNOWLEDMENTS

This work was funded by the ERDF through the Programme COMPETE and by the Portuguese Government through FCT – Project reference: UTA-Est/MAI/0007/2009.

## 2. REFERENCES

R. Aldecoa and I. Marín, 2011. “Deciphering network community structure by surprise”. PloS one, 6(9), e24195. doi:10.1371/journal.pone.0024195

Cravino, N., J. Devezas, and Á. Figueira (2012). Using the Overlapping Community Structure of a Network of Tags to Improve Text Clustering. In Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT 2012), Milwaukee, WI, USA.

Cutting et al, 1992. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceeding of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 318-329

Deerwester et al, 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391-407.

Devezas, J., and Á. Figueira (2012). Finding Language-Independent Contextual Supernodes on Coreference Networks. In IAENG International Journal of Computer Science, 39(2), pp.200-207.

Devezas, J., and Á. Figueira. The Community Structure of a Multidimensional Network of News Clips. IJWBC: Special issue on Community Structure in Complex Networks (accepted for publication)

Devezas, J., H. Alves, and Á. Figueira (2012). Creating News Context From a Folksonomy of Web Clipping. In Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2012 (IMECS 2012), Hong Kong.

Dumais, 1994, Latent semantic indexing (LSI):TREC-3 report. In Proceedings of the Third REtrieval Conference (TREC-3), pages 219-230.

Duval and Hodgins, 2004. Making metadata goes away: Hiding everything but the benefits. Keynote address at DC-2004, Shanghai, China, October.

Figueira et al, 2007. Interaction Visualization in Web-Based Learning using iGraphs. Hypertext’2007, 18th ACM Conference On Hypertext and Hypermedia. Manchester, UK. September, 2007.

G. Csárdi and T. Nepusz, “The igraph software package for complex network research,” InterJournal Complex Systems, vol. 1695, no. 1695, 2006. [Online]. Available: http://mycite.omikk.bme.hu/doc/14978.pdf

Gibson et al, 1998. Inferring Web communities from link topology. Proc. 9th ACM Conference on Hypertext and Hypermedia, 1998.

Harman, 1992, Relevance feedback revisited. In Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1-10.

J. Tang. Improved K-means Clustering Algorithm Based on User Tag. Journal of Convergence Information Technology, 5(10):124–130, 2010.

J. Xie, B. Szymanski, and X. Liu. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. Arxiv preprint arXiv:1109.5720, 2011.

Kleinberg, 1997. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997.

Kleinberg, 2000, Navigation in a Small World. Nature 406(2000), 845.

Kleinberg, 2000, The small-world phenomenon: An algorithmic perspective. Proc. 32nd ACM Symposium on Theory of Computing, 2000. Also appears as Cornell Computer Science Technical Report 99-1776 (October 1999).

L. Hubert and P. Arabie. Comparing partitions. Journal of Classification, 2(1):193–218, 1985.

L. Morey and A. Agresti. The measurement of classification agreement: an adjustment to the Rand statistic for chance agreement. Educational and Psychological Measurement, 44(1):33–37, 1984.

M. Ares, J. Parapar, and A. Barreiro. Improving Text Clustering with Social Tagging. In Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM 2011), pages 430–433, Barcelona, Spain, 2011.

M. Bastian, S. Heymann, and M. Jacomy, “Gephi: An open source software for exploring and manipulating networks,” in International AAAI Conference on Weblogs and Social Media, 2009, pp. 361–362. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/download/154/1009

R Development Core Team, “R: A language and environment for statistical computing,” in R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2011. [Online]. Available: http://www.r-project.org

Robertson and Sparc, 1976. Relevance weighting of search terms. Journal of American Society for Information Science, 27(3):129-146.

Rocchio, 1971, The SMART Retrieval System Experiments in Automatic Document Processing, Chapter on Relevance Feedback in Information Retrieval, pages 313-323. Prentice Hall.

Salton and Buckley, 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513-523.

Salton et al, 1975. A vector-space model for automatic indexing. Communications of the ACM, 18(11):613-620.

Scott, 2000, Social Network Analysis. Sage Publications Ltd; 2nd edition (March 2000). ISBN-13: 978-0761963394

Singhal, 1997, Term Weighting Revisited. PhD thesis, Cornell University.

Spink, 1995, Term relevance feedback and mediated database searching: Implications for information retrieval practice and systems design. Information Processing and Manegement, 31(2): 161-171.

Steinbach et al, 2000. A comparison of document clustering techniques. In Proceeding of Knowledge Data and Discovery (KDD). Workshop on Text Mining.

M. Steyvers and T. Griffiths, 2007. “Probabilistic Topic Models”. Handbook of Latent Semantic Analysis, 427(7), 424–440.

L. Tang, X. Wang and H. Liu, 2011. “Community detection via heterogeneous interaction analysis”. Data Mining and Knowledge Discovery. doi:10.1007/s10618-011-0231-0

U. Brandes, M. Eiglsperger, I. Herman, and M. Himsolt, “GraphML progress report structural layer proposal,” Graph Drawing, pp.501–512, 2002. [Online]. Available: http://www.springerlink.com/index/W6GU6JURTNWEF4MC.pdf

V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, p. P10008, 2008. [Online]. Available: http://iopscience.iop.org/1742-5468/2008/10/P10008

W. M. Rand. Objective Criteria for the Evaluation of Methods Clustering. Journal of the American Statistical Association, 66(336):846–850, 1971.

Willett, 1990, Document clustering using an inverted file approach. Journal of Information Science, 2:223-231.

X. Ji, W. Xu, and S. Zhu. Document clustering with prior knowledge. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 405–412, 2006.

Zhao and Karypis, 2002. Evaluations of algorithms for obtaining hierarchical clustering solutions. In Proceedings of the 2002 ACM International Conference on Information and Knowledge Management (ACM-CIKM), Washington D. C.

## GLOSSARY

**BCA** A hierarchical agglomerative clustering algorithm used to partition a set of observations, usually defined through the Vector Space Model, into *k* groups. It improves over *k*-means by choosing better initial cluster centroids.

**HITS** Hyperlink-Induced Topic Search is a ranking methodology used in link analysis to assign a hub and authority score to each node in a network. High hub scores indicate nodes with several connections to authorities, while high authority scores indicate nodes with several connections from hubs, so an authority is a highly cited node, while a hub is a node that cites a large number of authorities.

***k*-means** A clustering algorithm to partition a set of observations, usually defined through the Vector Space Model, into *k* groups. It uses *k* centroids, usually randomly assigned, assigns observations to its closest centroid, recalculates each clusters centroid and repeats, iterating until convergence.

Breadcrumbs: we start from the word vectors weighted using TF-IDF and extend this by integrating additional information from tags and comments.

**LSI**

Latent Semantic Indexing is a methodology that uses Singular Value Decomposition to cluster documents based on their semantic similarity.

**SLPA**

An overlapping community detection algorithm that identifies graph clusters based on label propagation, using labels to calculate the probabilities of membership for the nodes they represent.

Breadcrumbs: we use this algorithm to identify groups of related tags, a kind of social features.

**SVD**

Singular Value Decomposition is the factorization of a real or complex matrix that can be useful in signal processing or statistics.

Breadcrumbs: we use this technique for Latent Semantic Indexing as well as some of our community detection algorithms.

**TF-IDF**

Standing for Term Frequency / Inverse Document Frequency, this is a metric commonly used in information retrieval to measure the relevance of a term within a document, given a corpus of documents. It takes into account the number of times a term appears as well as how unique it is to the document. If a term appears many times in a document, but is rare in the whole collection, then it is highly relevant. On the other hand, if a term appears frequently in a document as well as the whole corpus, it loses relevance.

Breadcrumbs: we use this metric to assign weights to the terms of news clips, as well as their tags.

**VSM**

The Vector Space Model is an algebraic model used to represent documents through their features, usually the terms, enabling document comparison and clustering by using metrics such as the euclidean distance or the cosine similarity.

Breadcrumbs: we use the Vector Space Model in Breadcrumbs to describe our documents along with their tags and comments.