Creating News Context From a Folksonomy of Web Clipping

José Devezas, Henrique Alves, and Álvaro Figueira

***Abstract*—We propose a method for creating news context by taking advantage of a folksonomy of web clipping based on online news. We experiment with an ontology-based named entity recognition process and study two different ways of modeling the relationships induced by the coreference of named entities on news clips. We try to establish a context by identifying the community structure for a clip-centric network and for an entity-centric network, based on a small test set from the Breadcrumbs system. Finally, we compare both models, based on the detected news communities, and show the advantages of each network specification.**

***Index Terms*—named-entity-recognition, semantic-analytics, relationship-extraction, community-detection**

I. Introduction

Today's conjuncture of digital and social media is merging the roles of readers and providers. While readers are increasingly participating in the news media sites commenting news and participating in discussion forums, journalists are eager to get feedback from their readers, which will eventually lead them to enhance a story or even to a new story. It is fair to say that, today, the future of news depends on harnessing the participation of readers in the global process of production and consumption of news.

In this article we use the "Breadcrumbs" system [1] 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". Breadcrumbs uses social web tools to gather the opinions of readers, and creates a semantically organized model of the readers' opinions. In particular, Breadcrumbs focuses on: collecting news fragments; organizing those fragments automatically and aggregating the fragments across readers.

In this paper we take a step forward and present a method for inferring relationships between readers, and for inferring relationships between news.

A. 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. Breadcrumbs extends this by allowing people to keep track of news at a fine-grained detail level. Breadcrumbs lets readers select news stories fragments from any news site or blog using a dedicated, and web based tool. By collecting these fragments, or "clips", readers are automatically feeding their own "Personal Digital Library" (PDL). In addition each clip can be annotated with tags and/or comments.

While each PDL represents the individual perspective of a reader, 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, Breadcrumbs uses automatic mechanisms that classify the news fragments based on their content and semantic proximity [2], [3]. As for the PDL aggregation, we focus on text mining and social classification methods [4] that potentially identify implicit links or relationships between fragments based not only on text similarity, but also on the tags and comments assigned by the users.

The relationship inferencing system already implements a set of rules capable of establishing strong and weak ties between clips. We propose an approach based on semantic analytics, that aims to extend the system by taking advantage of named entity recognition in order to identify relevant knowledge, such as people, places or even dates. This will allow us to establish new relationships between clips, providing insights into the Five Ws — who, where, when, what and why — of the news communities in our corpus.

II. Breadcrumbs: A Folksonomy of Web Clipping

Breadcrumbs is an ongoing research project that aims at creating a social network based on the relations established by collections of fragments taken from online news. On one hand, the system enables registered users to do news clipping, that is, collect and store online news text fragments, or clips, which are for some reason of interest to them. On the other hand, it allows for smart classification of clip collections and eases the process of obtaining more related news through a novel browsing system based on identified relations of similar news clips from other users.

A clipping task involves selecting a part of the news text, assigning it a title, eventually personal comments and at least one tag, hence creating a folksonomy [5] from online news clips. In order to do that, Breadcrumbs allows seamless web browsing to online news providers through its own proxy, which injects a piece of JavaScript code into the page. This injected code enables the user to easily obtain the fragments of news that he wants to collect, without the hassle of having to install software like a browser specific extension.

These clips are then stored and organized for future reference in the user's PDL: the system's graphical user

Manuscript received December 17, 2011; revised January 17, 2012. This work was supported in part by FCT Project Breadcrumbs UTA-Est/MAI/0007/2009.

J. Devezas, H. Alves and A. Figueira are with CRACS/INESC TEC, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 1021/1055, 4169-007 Porto, Portugal e-mail: jld@dcc.fc.up.pt, halves@alunos.dcc.fc.up.pt, arf@dcc.fc.up.pt.

interface for clip collection management. The PDL allows users to visually organize their clips in two ways: manually or automatically. For instance, users may ask Breadcrumbs to organize their collection of clips for them. In this case, the system performs an automatic social classification based on the clips content and their respective tags, effectively computing clusters of semantically related news clips, thus providing the user an organized collection. From that point forth he may proceed with his own manual adjustments to the classification, or simply accept it. Users are able to organize their collection of clips in a natural way, as they would do with a collection of interesting papers in a real desktop, grouping clips that, in their point of view, are related. From this natural organization, specific structures will arise like piles, matrices or even circles of clips. Those types of structures are detected by the PDL which informs the system of that particular user's choices to organize his clips, providing useful information for future uses of the automatic social classification feature and for the identification of relations between clips.

As more users take advantage of this system to manage and maintain their collections of news clips, chances are that different users might have taken fragments of the same original news. Some of those fragments, taken from the same source, will probably even overlap. From these situations, new kinds of relations of common interests arise, enriching the possibilities to find more content of potential interest. For instance a user might want to get more stories similar to one of his clips. To do that, he may browse from his PDL to another user's PDL but only to access a subset of it: the group or cluster to which the related clip belongs, effectively providing more related news from possibly different sources. This kind of browsing through PDL's requires the identification of pertinent relations between collections, groupings and particular clips.

We take advantage of this system and use a small data set with over 250 news clips as a test set for the model we propose here, which, in turn, could be integrated into the Breadcrumbs system as another interesting and meaningful way to further identify relationships between news clips, capable of enriching the semantics of the system.

III. Ontology-Based Named Entity Recognition

Named entity recognition is a problem that is commonly solved by using linguistic grammar-based techniques [6] as well as statistical modelling methods, such as conditional random fields [7]. However, for this specific problem, our system uses a different approach, based on the semantic web. At the cost of having to preselect a good set of classes from one of the available ontologies, we are able to obtain a set of entity lists that have been, and continue to be, curated over time by the online community. Additionally, we have access to translations of these entity lists in more than one language, making it possible to find matches in web clips independently of the language, and resolving each match to the corresponding resource URI. This also allows us to establish language-independent relationships between clips or entities, which results in richer networks capable of providing better insights into the context of web clips.

DBpedia [8] is an openly available and highly curated and complete data set that provides semantically structured information based on Wikipedia, as well as a public SPARQL endpoint to query the data set. Based on previous work by Devezas et al. [9], where they studied a personality coreference network from a news stories photo collection, we picked some of DBpedia's ontology subclasses of *Person*, focusing on the topics of politics, sports and finance, and additionally considering some of the art-related topics. The list of entities for each of the selected subclasses serves as the knowledge base for identifying people in news clips and answering the question "Who?". Similarly, we select a set of subclasses of *Place* to helps us find an answer to the question "Where?", and finally we suggest that the question "When?" should be answered by taking advantage of the various date properties, available in relevant resources such as DBpedia's *Event*, or by using the YAGO's knowledge base [10], [11].

IV. Inferring Context From the Analysis of Named Entity Coreference Networks

As the study by Devezas et al. [9] points out, the community structure of a personality coreference network can provide insightful information about the context of news stories. They used the short photo descriptions, a fragment of text previously selected from the full news story where the photo appeared, to identify personalities and build the coreference network. Then, by running a community detection algorithm, they identified the network's community structure and were able to assign keywords to each community by aggregating and analysing the photo descriptions where the personalities in the community were mentioned.

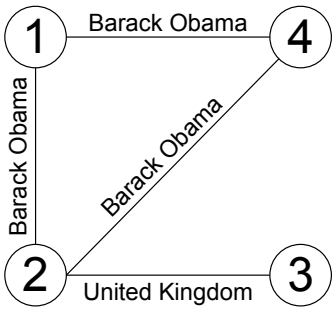
We believe that community detection methodologies can similarly be applied to our data set, for the analysis of a more complete named entity network, where people, places and dates can all be connected if coreferenced in a web clip. Based on this network, we can then find communities (clusters induced by the clipping behavior of people), capable of providing insights into the context of our corpus, as an attempt to answer the questions "What?" and "Why?", by emphasizing the highly related and densely connected groups of entities that were identified and validated with the help of the semantic web.

Additionally, we are interested in experimenting with a clip-centric network, where relationships between clips are established by the coreference of an entity in a pair of distinct clips, as opposed to an entity-centric network, where relationships are established by the coreference of a pair of distinct entities in a single clip. The clip-centric model has the advantage of enabling the direct mapping of results into clips instead of entities, but given the difference in paradigm it is uncertain whether or not it will produce similar groups of information. So we test both models.

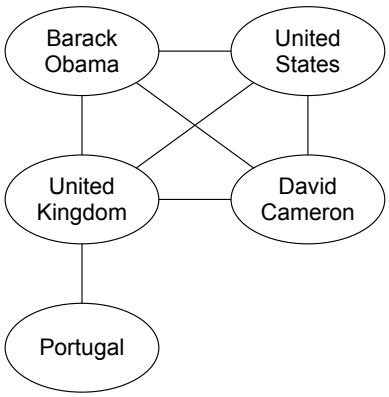
TABLE I illustrates a possible result from the ontology-based named entity recognition process described in the previous section. Given four clips, our system identified four named entities for clip 1, one for clip 2, two for clip 3 and one for clip 4. For the purpose of referencing each entity, we also include the corresponding label in English. Keep however in mind that named entities can be identified in any of the languages available in the knowledge base and are then resolved to their corresponding URI before building the graphs. Fig. 1 depicts two alternate methods for modeling the results in TABLE I. In Fig. 1a, we show

Clip ID	Entity URI	Entity Label
1	http://dbpedia.org/resource/Barack_Obama	Barack Obama
1	http://dbpedia.org/resource/United_States	United States
1	http://dbpedia.org/resource/United_Kingdom	United Kingdom
1	http://dbpedia.org/resource/David_Cameron	David Cameron
2	http://dbpedia.org/resource/Barack_Obama	Barack Obama
3	http://dbpedia.org/resource/Portugal	Portugal
3	http://dbpedia.org/resource/United_Kingdom	United Kingdom
4	http://dbpedia.org/resource/Barack_Obama	Barack Obama

TABLE I: Example of named entities identified in four clips.



(a) Clip-centric network.



(b) Entity-centric network.

Fig. 1: Two types of networks to model named entity coreferencing in web clips.

a clip-centric network model, where clips 1, 2 and 4 are connected because they all mention "Barack Obama" and clips 2 and 3 are connected because they both mention "United Kingdom". As you can see from this theoretical example, some information is already lost, since there is no reference to "United States", "David Cameron" or "Portugal". On the other hand, the strongest relations between clips are in fact imposed by "Barack Obama" and "United Kingdom". In Fig. 1b, we show an entity-centric network model, where "Barack Obama", "United States", "United Kingdom" and "David Cameron" are all connected because they are coreferenced in clip 1 and "United Kingdom" and "Portugal" are both connected because they are coreferenced in clip 3. While this model captures all of the information available, it also requires some sort of index structure to obtain all the clips where each entity was mentioned, that will work as a translation mechanism from named entity to clip, after identifying the community structure.

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*, *NaturalPlace* and *HistoricPlace* — and *Person* subclasses — *Politician*, *OfficeHolder*, *Athlete*, *Cleric*, *Scientist*, *Model*, *Criminal* and *Judge*. An early experiment with the classes *Artist*, *Band* and *Organisation* resulted in a large set of misidentified entities, corresponding to unusual names, that would match against parts of the sentence that did not represent real named entities. This is a clear indication of the absence of tradicional natural language processing methodologies, emphasizing the importance of a grammatical analysis in order to identify the phrase structure. We have abstained from following this path from the beginning, as these methodologies are usually language-dependent and we were interested in experimenting with language-independent techniques based on the semantic web.

A single news clip will ideally be pertinent to its creator and will possibly contain some of the most relevant information of the news story. 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. We pre-process the data from the Breadcrumbs system using the R Project [12] and the igraph package [13], transforming the clip–entity dictionary into two GraphML [14] files, one for each network model. Using Gephi [15], we do an exploratory visual analysis of both networks, calculating the eigenvector centrality for every node in the graph, with 100 iterations, and identifying their community structure using the modularity-based methology by Blondel et al. [16]. Next, we describe the results of this analysis, presenting additional data about the communities, and evaluating our attempt to create news context from a folksonomy of web clipping.

A. Using a Clip Network Induced by the Coreference of Similar Named Entities Across Distinct Clips

Fig. 2 depicts the community structure of the clip-centric network model for the 259 clips in the Breadcrumbs database. We have established a connection between a pair of clips, whenever they both mentioned the same entity (in any

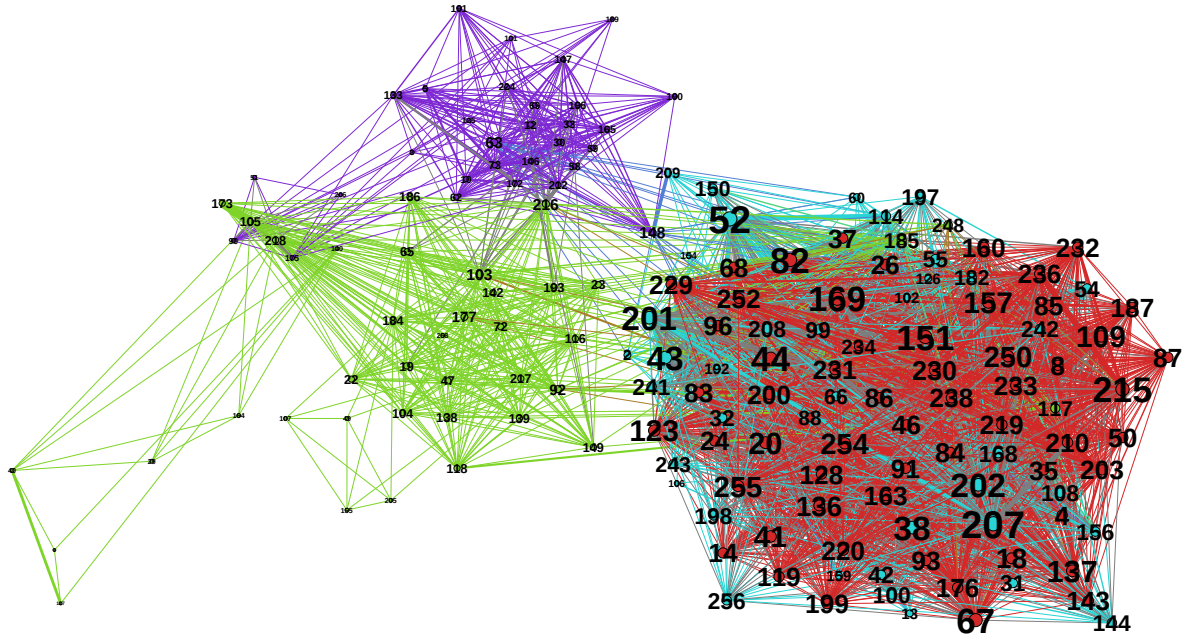


Fig. 2: Community structure for the largest component of the clip network.

language available in DBpedia). This resulted in a network with 175 nodes, connected by 3,333 edges, with a density of 21.89% and a diameter of 5. We have analyzed the largest component of the graph, identifying four large communities, which are further described in TABLE II.

B. Using a Named Entity Network Induced by the Coreference of Distinct Named Entities in the Same Clip

Similarly, Fig. 3 depicts the community structure of the entity-centric network model for the same 259 clips. In this model, we have established a connection between a pair of entities, whenever they were mentioned together (coreferenced) in a clip. Since the entities had been previously resolved to their corresponding URI, we could say that we are trying to establish a language-independent context. 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 III.

V. Comparing The Models

We compare the models by analyzing the most prominent communities in each network, as an attempt to determine the most informational model. 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 assign keywords (see TABLE II) that describe the content in each clip. Since some of the users had clipped the same fragment of the news story, we can find the same exact keywords for two consecutive clips. The fact that they have 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 5 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.

As we can see from TABLE II, community 0 establishes a context for the economic crisis in the United States of America, where tax raising is discussed in diverse situations. Communities 3 and 7 establish a context for the economic crisis in Europe. Visually, these two communities almost overlap, which indicates a strong connection between them. Even though they are topically identical, each one covers different aspects of the European economic crisis — community 3 refers to Japan's interest in buying European bonds, while community 7 focuses on the Euro and other currency-related affairs, such as bank recapitalization. Finally, community 6 establishes a context for the Libyan revolution, part of the Arab Spring, a wave of demonstrations and protests in the Arab world.

TABLE III shows the top 5 nodes for the three main communities identified in the entity network. As we can see, one of the nodes is labeled "The President" and was wrongly identified during the named entity recognition phase. This happened because this is a common expression on news stories and it can also be recognized as a mountain peak in Canada, which was part of the DBpedia's *NaturalPlace* entity set that we used. The community structure of this

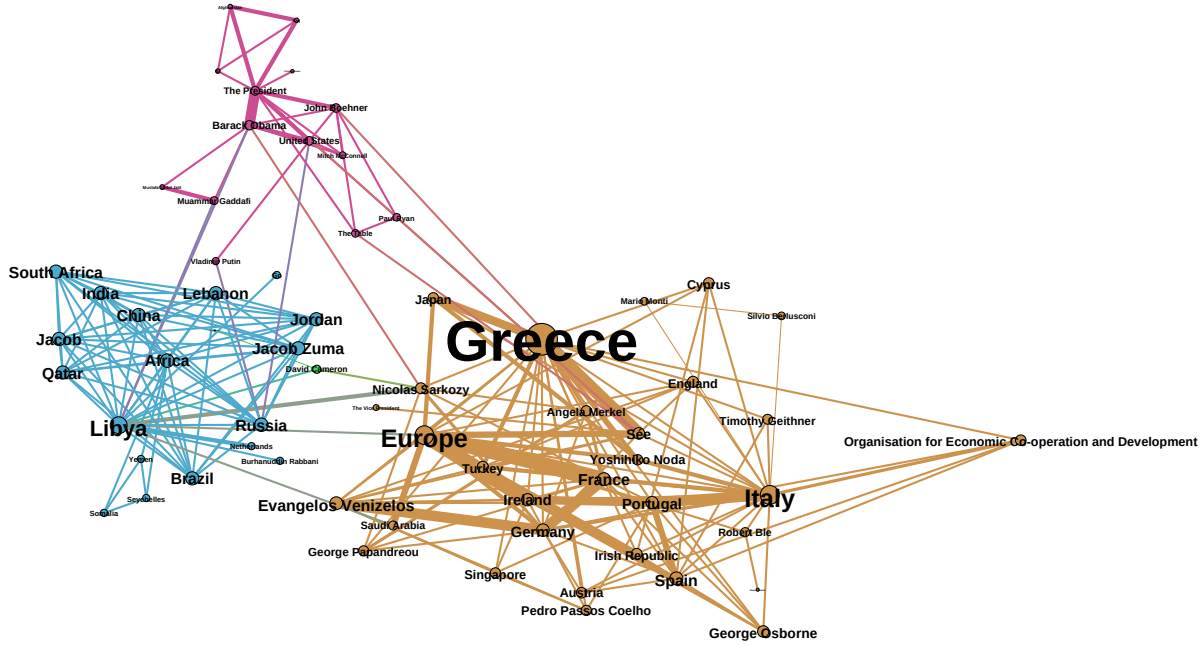


Fig. 3: Community structure for the largest component of the entity network.

network clearly separates the topics of the news corpus, but also identifies new coreferences, such as Barack Obama and Muammar Gaddafi. Community 5 aggregates entities about the European economic crisis, community 7 aggregates United State affairs, showing a weaker but still present connection to the Arab Spring, which is in turn visible in community 9. By looking at Fig. 3, the most relevant entities are immediately recognizable. We can see Greece as a central reference and the Organization for Economic Co-operation and Development as an indicator of the news community topic. Additionally, we notice that this is a visualization-friendly model, as there are fewer nodes, a more illustrative community structure and weighted edges that depict the strength of ties.

As the communities evolve, and our corpus of news clips grows, it is possible that the topic of each community becomes more prominent, further emphasizing the borders around communities. On the other hand, topics might evolve into several subtopics, in which case communities will split into smaller communities, but it can also happen that two topics become more connected over time, in which case communities will merge into a larger community [17]. Either way, as the corpus grows and evolves, the insights into the context of each news community will be improved.

Community ID	EVC	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

TABLE III: Analysis of the main communities identified for the entity network (EVC stands for eigenvector centrality).

VI. Conclusions and Future Work

We have extracted and studied the relationships between news clips based on named entities and proposed a method for creating news context using the Breadcrumbs system as a folksonomy of web clipping. We explored two different ways of modeling the underlying relationships found through a clip–entity dictionary. We briefly compared the two models and found them both to be viable in the task of describing 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. 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

Community ID	EVC	User ID	Clip ID	Keywords
0	0.080254	6	148	USA; Barack Obama; Economy; Crisis; Billionaire; Tax
0	0.080254	2	63	USA; Barack Obama; Economy; Crisis; Billionaire; Tax
0	0.025265	6	212	USA; Barack Obama; Economy; Crisis; Tax
0	0.025265	4	73	USA; Barack Obama; Economy; Crisis; Tax
0	0.025265	2	12	USA; Barack Obama; Economy; Crisis; Tax; Health Insurance
3	1.000000	4	52	Europe; Economy; Crisis; Summary; Italy; Greece; UK
3	1.000000	6	207	Europe; Economy; Crisis; Summary; Italy; Greece; UK
3	0.986015	6	202	Europe; Economy; Crisis; Italy; Rating; IMF; Greece; Japan; Bonds
3	0.986015	6	201	Europe; Economy; Crisis; Italy; Rating; IMF; Greece; Japan; Bonds
3	0.986015	4	43	Europe; Economy; Crisis; Italy; Rating; IMF; Greece; Japan; Bonds
6	0.303381	6	185	NATO; Netherlands; Libya
6	0.303381	3	117	NATO; Netherlands; Libya
6	0.188308	1	248	USA; New York; Traffic Fines; Scandal
6	0.082941	4	103	United Nations; Libya; Moammar Gadhafi; Mustafa Abdul-Jalil; Barack Obama
6	0.082941	6	216	United Nations; Libya; Moammar Gadhafi; Mustafa Abdul-Jalil; Barack Obama
7	0.996523	4	82	Europe; Crisis; Summary; Greece; Ireland; Portugal; IMF; Italy; Spain
7	0.996523	6	215	Europe; Crisis; Summary; Greece; Ireland; Portugal; IMF; Italy; Spain
7	0.996523	3	44	Europe; Crisis; Greece; Ireland; Portugal; Italy; Spain; Bank Recapitalization
7	0.993462	2	67	Europe; Crisis; Rating; Italy; Greece; Spain; Ireland; Cyprus; Euro
7	0.993462	6	151	Europe; Crisis; Rating; Italy; Greece; Spain; Ireland; Cyprus; Euro

TABLE II: Analysis of the main communities identified for the clip network (EVC stands for eigenvector centrality).

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.

As future work, we would like to further investigate the contents of each community, in an attempt to provide a better evaluation scheme for our models. We would also like to experiment with a larger corpus of news clips, allowing us to create context for a wider range of news topics. Finally, we would like to improve the ontology-based named entity recognition process and take advantage of the semantic web to make inferences on the discovered knowledge, providing additional contextual details to the user, and even suggest him additional news sources.

REFERENCES

[1] Álvaro Figueira et al., "Breadcrumbs: A social network based on the relations established by collections of fragments taken from online news," *Retrieved December 5, 2011, from http://breadcrumbs.up.pt*.

[2] G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[4] D. Gibson, J. Kleinberg, and P. Raghavan, "Inferring web communities from link topology," in *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space Structure in Hypermedia Systems*. New York, New York, USA: ACM, 1998, pp. 225–234.

[5] T. Vander Wal, "Folksonomy Coinage and Definition," *Retrieved December 5, 2011, from http://vanderwal.net/folksonomy.html*.

[6] A. Mikheev, M. Moens, and C. Grover, "Named entity recognition without gazetteers," in *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1999, pp. 1–8.

[7] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, vol. 4. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 188–191.

[8] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia - A crystallization point for the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154–165, Sep. 2009. [Online]. Available: http://dbpedia.org/About

[9] J. Devezas, F. Coelho, S. Nunes, and C. Ribeiro, "Studying a Personality Coreference Network in a News Stories Photo Collection," in *Proceedings of the 34th European Conference on Information Retrieval (ECIR 2012)*, 2012.

[10] F. M. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web - WWW '07*. New York, New York, USA: ACM Press, 2007, p. 697. [Online]. Available: http://www.mpi-inf.mpg.de/yago-naga/yago/

[11] J. Hoffart, F. Suchanek, K. Berberich, E. Lewis-Kelham, G. De Melo, and G. Weikum, "YAGO2: exploring and querying world knowledge in time, space, context, and many languages," in *Proceedings of the 20th International World Wide Web Conference (WWW 2011)*. ACM, 2011, pp. 229–232. [Online]. Available: http://www.mpi-inf.mpg.de/yago-naga/yago/

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] D. Greene, D. Doyle, and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *2010 International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2010, pp. 176–183.