Finding Language-Independent Contextual Supernodes on Coreference Networks

José Devezas 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, describing two alternate implementation approaches, and we 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. Reference Work

While reaching an agreement on the formal definition of context is not an easy task, some work has been done in an attempt to characterize and use it. Dey and Abowd [5] have surveyed the use of context in interactive applications, focusing on context-aware applications. Based on the analysis of other work done on this topic, where other researchers had attempted to define context, they proposed their own unifying definition of context as "any information that can be used to characterize the situation of an entity". Yeung et al. [6] have studied the context of tags in folksonomic systems, using community detection methodologies to identify groups of semantically similar tags in order to facilitate disambiguation.

A community is by itself the definition of a social context [7]. Mathematically, a community in a graph has been defined as a densely connected subgraph. This means that nodes in a community are more frequently connected to each other than they are with the remaining nodes in the network.

Manuscript received May 10, 2012; revised May 14, 2012. This work is financed by the ERDF — European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT — Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project UTA-Est/MAI/0007/2009.

J. Devezas 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, arf@dcc.fc.up.pt.

There are several methods for identifying these special contextual groups. Community detection methodologies can be categorized according to the type of network (weighted or unweighted, connected or disconnected, directed or undirected, multidimensional/multimodal), or according to the resulting partitions (hierarchical, overlapping or non-overlapping). State of the art methodologies for the identification of community structure include the Louvain method [8], Surprise maximization [9], or the Speaker-listener Label Propagation Algorithm (SLPA) [10]. The Louvain method is a multilevel aggregation algorithm based on the local optimization of the modularity score — a metric proposed by Newman and Girvan [11] to measure the quality of a network partition. Given the resolution limits of the modularity score, identified by Fortunato and Barthélemy [12], a new metric called Surprise has been introduced by Aldecoa and Marín [9] to globally measure the quality of a partition taking into account the number of nodes and edges in the communities. The maximization of either quality metric results in a disjoin partition of the network. However, in real-world networks, it is frequently relevant to find the overlapping community structure, where a node might belong to more than one community simultaneously. SLPA is one of the state of the art algorithms that enables the identification of a fuzzy partition of overlapping communities through the propagation of node labels. These labels are kept in memory so they can, later on, be used to calculate probabilities of membership for the nodes they represent. Tang et al. [13] have also worked on community detection, contributing on a different aspect of the area. They suggested a unified view for methodologies that are based on the input of an adjacency matrix, which made it possible to devise an integration strategy for the identification of community structure in multidimensional networks using the already available algorithms. They proposed that the integration operation could be done by combining the resulting matrices for each dimension, in four possible steps of the detection procedure: *(i)* network/adjacency matrix integration, *(ii)* utility matrix integration, *(iii)* structural features matrix integration, and *(iv)* partition matrix integration.

The work we present here is an extended version of our previous paper, entitled “Creating News Context from a Folksonomy of Web Clipping” [14], published in the IMECS 2012 proceedings. In this version, we have added detailed insights on the implementation of our named entity recognition system, as well as an elaboration on the reference work and an extended discussion on the topical coherence of the obtained contexts.

III. 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 [15] 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 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.

A. System Architecture

Fig. 1 depicts a simplistic overview of the Breadcrumbs system architecture. The illustrated components can be divided into *(i)* a front end, corresponding to the components that directly interact with the user, *(ii)* a back end, comprised of the intelligent subsystems and service provider components, *(iii)* a middle layer implemented using a web service architecture based on a REST API, and *(iv)* a storage layer, corresponding to an internal repository API, which connects to a relational database such as MySQL or other NoSQL services such as Couchbase. In the front end layer, we can find the Clipper, corresponding to a small dashboard injected in every visited web page, through the back end Proxy service, to allow the user to create and store a web clip in the Data Repository. Users can use

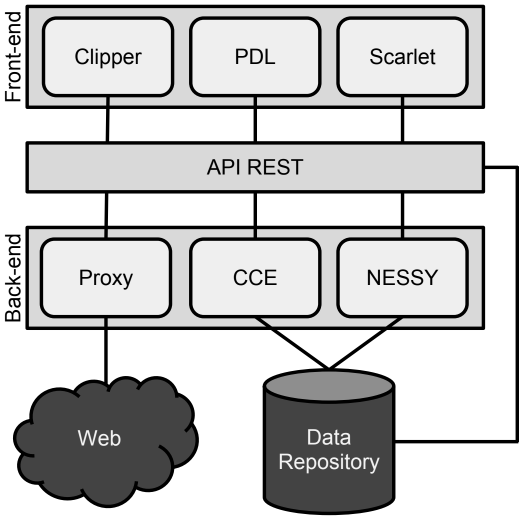


Fig. 1: Breadcrumbs system architecture overview.

the PDL module to view their collected clips, eventually organizing them into groups, doing so manually or with the aid of the Classification and Clustering Engine (CCE). The final intelligent subsystem in the Breadcrumbs architecture is the Named Entity Sensory SYstem (NESSY), which we describe in detail in this paper. NESSY acts as a named entity recognition system, identifying people, places and dates, while inferring relationships between these entities and their corresponding clips, which results in a network with community structure. The generated networks can then be visualized by using the Scarlet interface, either in the form of a multidimensional network displayed using a force-directed algorithm [16], [17], or in the form of a multiresolution network map, using the *gvmap* algorithm by Gansner et al. [18].

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.

IV. Ontology-Based Named Entity Recognition

Named entity recognition is a problem that is commonly solved by using linguistic grammar-based techniques [19] as well as statistical modelling methods, such as conditional random fields [20]. 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 [21] 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. [22], 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 [23], [24].

A. Implementation Methodologies

In this section, we describe the two methods we used to implement the ontology-based named entity recognition process. First, as a prove of concept, we followed a naive approach based on regular expressions. We began by detecting the clip's language, using a Java language detection library[1] capable of analyzing small fragments of text. The next step was to retrieve the labels for the previously selected collection of DBpedia entities using the same language as the one identified for the web clip. To do this, we used a three step caching system that first tried to find a collection of entities loaded in memory. On a missing hit, the local database was queried. Whenever this yielded no results, DBpedia's SPARQL endpoint was used. As expected, each successful hit was back-propagated, to the faster caching system. After loading the entities into memory, a regular expression was precompiled for each entry to find all matches of the entity's label surrounded by a word separator to each text fragment contained in a clip.

Even though this first method was convenient during the research process, to quickly obtain the entities and thus the network models induced by their co-occurrence, a better implementation was needed for the production system. During the optimization and scalability development phases, we researched possible data structures and algorithms to improve the matching procedure. This particular type of problem could easily fit the category of string searching algorithms using a finite set of patterns, that is, using a dictionary of predefined strings as the set of patterns to be identified in a text. Several alternatives were available for this type of procedure. We opted for the same implementation as the Unix tool *fgrep*, a widely available and documented alternative. The algorithm we used, Aho-Corasick [25], is based on a finite state machine constructed using a custom trie (or prefix tree) with additional links between the internal nodes. The Aho-Corasick algorithm has a time complexity of $O(m + n)$ for a text of length m, containing n pattern matches. Predictably, using the Aho-Corasick algorithm instead of the naive approach based on regular expressions introduced a significant speedup to our system, increasing performance by two orders of magnitude.

Further improvements could still be done, such as replacing the Aho-Corasick algorithm by the Commentz-

[1] http://code.google.com/p/language-detection/

Walter [26] alternative, which has worst case search time complexity $O(mL)$, where L is the maximum pattern length. Even though the worst case scenario for these algorithms doesn't favor the Commentz-Walter alternative, this algorithm has been reported to achieve better performance in practise. Additionally, since we have access to a document database, where we can store serializable Java objects, another possible improvement would be to periodically create the trie structure, caching it for future use, thus reducing the overhead introduced by the initialization process.

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

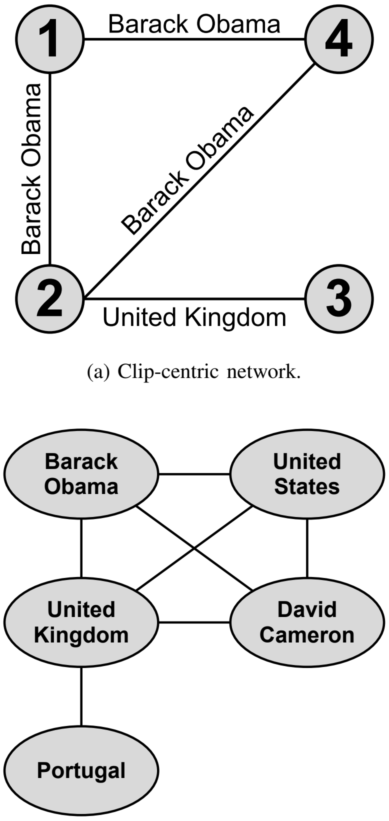
As the study by Devezas et al. [22] points out, the community structure of a personality coreference network can provide insightful information about the context of news stories. They used the short photo descriptions, a fragment of text previously selected from the full news story where the photo appeared, 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.

A. Describing the Theoretical Models

TABLE I illustrates a possible result from the ontology-based named entity recognition process described in the previous section. Given four clips in two different languages (English and Portuguese), our system identified four named entities for clip 1, one for clip 2, two for clip 3 and one for clip 4. Fig. 2 depicts two alternate methods for modeling the results in TABLE I. For the purpose of referencing each entity, we use 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. In Fig. 2a, we show 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" — notice that clip 2 contains an instance of "United Kingdom" in English, while clip 3 contains an instance of "United Kingdom" in Portuguese; nevertheless, our goal is to establish connections independently of the language. 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. 2b, 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.



(a) Clip-centric network.

(b) Entity-centric network.

Fig. 2: Two types of networks to model named entity coreferencing in web clips.

B. Experimenting with Real Data

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

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

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

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 [27] and the igraph package [28], transforming the clip–entity dictionary into two GraphML [29] files, one for each network model. Using Gephi [30], 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. [8]. 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.

1) Using a Clip Network Induced by the Coreference of Similar Named Entities Across Distinct Clips: Fig. 3 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 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.

2) Using a Named Entity Network Induced by the Coreference of Distinct Named Entities in the Same Clip: Similarly, Fig. 4 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.

VI. 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 five 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 do an analysis of the text, supported by the frequent term set, and then manually assign keywords to each of the five groups of nodes, in order to illustrate the topics of the communities (see TABLE II). Since some of the users had clipped the same fragment of the news story, we can find the same exact terms 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 five nodes of the entity-centric network, since we can 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 five 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

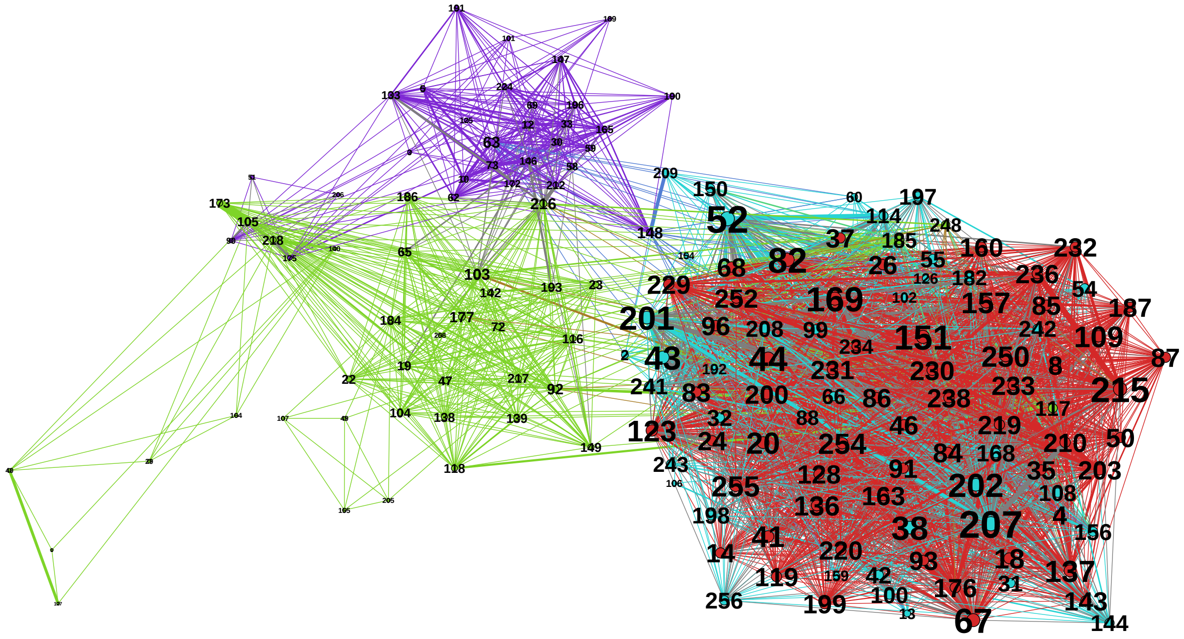


Fig. 3: Community structure for the largest component of the clip network.

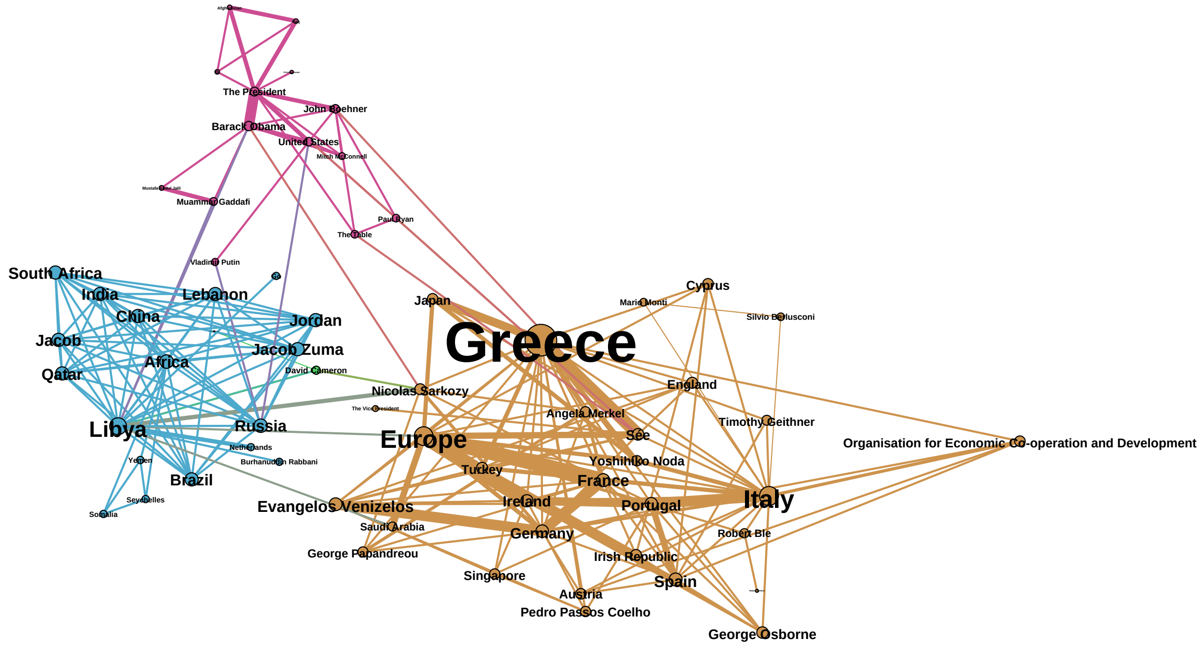


Fig. 4: Community structure for the largest component of the entity network.

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

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

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 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. 4, 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 [31]. Either way, as the corpus grows and evolves, the insights into the context of each news community will be improved.

VII. Conclusions

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 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.

VIII. Future Work

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] A. Figueira, P. Ribeiro, J. P. Leal, F. Zamith, E. Cunha, L. Francisco-Revilla, H. Ribeiro, A. Silva, M. Pinto, H. Alves, J. Devezas, M. Santos, and N. Cravino, "Breadcrumbs: A social network based on the relations established by collections of fragments taken from online news," *Retrieved January 19, 2012, from http://breadcrumbs.up.pt*, 2009.

[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] A. K. Dey and G. D. Abowd, "Towards a better understanding of context and context-awareness," in *Handheld and Ubiquitous Computing*, 1999, pp. 304–307. [Online]. Available: http://www.springerlink.com/index/PWPMM42N3KRR1F3A.pdf

[6] C. M. Au Yeung, N. Gibbins, and N. Shadbolt, "Contextualising tags in collaborative tagging systems," in *Proceedings of the 20th ACM conference on Hypertext and hypermedia.* ACM, 2009, pp. 251–260. [Online]. Available: http://dl.acm.org/citation.cfm?id=1557958

[7] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0370157309002841

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, Oct. 2008. [Online]. Available: http://iopscience.iop.org/1742-5468/2008/10/P10008http://arxiv.org/abs/0803.0476http://stacks.iop.org/1742-5468/2008/i=10/a=P10008?key=crossref.46968f6ec61eb8f907a760be1c5ace52

[9] R. Aldecoa and I. Marín, "Deciphering network community structure by surprise." *PloS one*, vol. 6, no. 9, p. e24195, Jan. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3164713\&tool=pmcentrez\&rendertype=abstract

[10] 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. [Online]. Available: http://arxiv.org/abs/1109.5720

[11] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 026113, 2004. [Online]. Available: http://pre.aps.org/abstract/PRE/v69/i2/e026113

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

Community ID	EVC	User ID	Clip ID	Frequent Terms	Topical Keywords
0	0.080254	6	148	tax; president; americans; buffet	USA; Tax; Billionaire; Economy; Crisis; Barack Obama; Health Insurance
0	0.080254	2	63	tax; president; americans; boehner	
0	0.025265	6	212	tax; president; earning; class	
0	0.025265	4	73	tax; president; earning; plans	
0	0.025265	2	12	tax; american; trillion; medicare	
3	1.000000	4	52	minutes; greek; austerity; tax	Europe; Economy; Crisis; Summary; Italy; Rating; Greece; IMF; Japan; Bonds
3	1.000000	6	207	minutes; greek; austerity; bank	
3	0.986015	6	202	european; markets; bonds; noda	
3	0.986015	6	201	european; markets; bonds; international	
3	0.986015	4	43	european; markets; bonds; monetary	
6	0.303381	6	185	nations; economic; libya; petroleum	United Nations; NATO; Netherlands; Libya; Moammar Gadhafi; Mustafa Abdul-Jalil; Barack Obama
6	0.303381	3	117	nations; economic; libya; oil	
6	0.188308	1	248	european; sarkozy; merkel; banking	
6	0.082941	4	103	gadhafi; war; libya; international	
6	0.082941	6	216	gadhafi; war; libya; democracy	
7	0.996523	4	82	greece; bonds; debt; bailout	Europe; Crisis; Summary; Greece; Ireland; Portugal; IMF; Italy; Spain; Bank Recapitalization; Cyprus; Euro
7	0.996523	6	215	greece; bonds; debt; euro	
7	0.996523	3	44	banks; sovereign; europe; governments	
7	0.993462	2	67	nations; euro; debt; crisis	
7	0.993462	6	151	nations; euro; debt; crisis	

[12] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, p. 36, 2007. [Online]. Available: http://www.pnas.org/content/104/1/36.short

[13] L. Tang, X. Wang, and H. Liu, "Community detection via heterogeneous interaction analysis," *Data Mining and Knowledge Discovery*, Aug. 2011. [Online]. Available: http://www.springerlink.com/index/10.1007/s10618-011-0231-0

[14] J. Devezas, H. Alves, and A. Figueira, "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*, 14-16 March, 2012, Hong Kong, pp. 446–451.

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

[16] T. Fruchterman and E. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/spe.4380211102/abstract

[17] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-Driven Documents." *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2301–9, Dec. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22034350

[18] E. Gansner, Y. Hu, and S. Kobourov, "GMap: Drawing Graphs as Maps," in *Graph Drawing*. Springer, 2010, pp. 405–407. [Online]. Available: http://www.springerlink.com/index/j711430403tux055.pdf

[19] 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.

[20] 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.

[21] 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

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

[23] 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/

[24] 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/

[25] A. Aho, "Efficient string matching: an aid to bibliographic search," *Communications of the ACM*, vol. 18, no. 6, pp. 333–340, Jun. 1975. [Online]. Available: http://portal.acm.org/citation.cfm?doid=360825.360855http://dl.acm.org/citation.cfm?id=360855

[26] B. Commentz-Walter, "A string matching algorithm fast on the average," in *Automata, Languages and Programming*, 1979, pp. 118–132. [Online]. Available: http://www.springerlink.com/index/FW657X6118785611.pdf

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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.