Studying a Personality Coreference Network in a News Stories Photo Collection

José Devezas[†], Filipe Coelho[‡§], Sérgio Nunes[‡§], and Cristina Ribeiro[‡§]

[†]Labs SAPO/UP & [§]INESC TEC &
[‡]DEI, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
{jld, filipe.coelho, ssn, mcr}@fe.up.pt

Abstract. We build and analyze a coreference network based on entities from photo descriptions, where nodes represent personalities and edges connect people mentioned in the same photo description. We identify and characterize the communities in this network and propose taking advantage of the context provided by community detection methodologies to improve text illustration and general search.

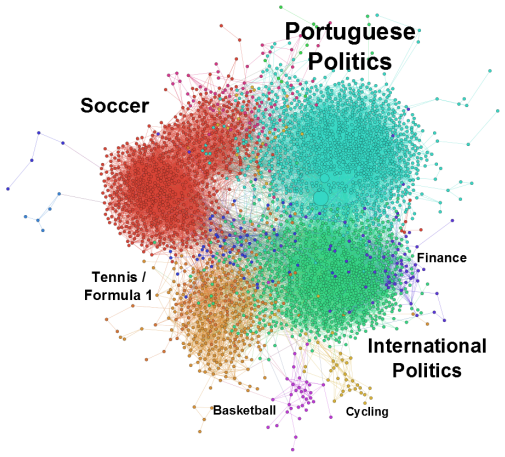
1 Introduction

Motivated by the idea of grouping photos into contextual groups or communities in order to improve their use in text illustration, we extract and experiment with entities from photo descriptions, building and analyzing a coreference network, where nodes represent personalities and edges connect people mentioned in the same photo description. We are interested in studying the community structure of this kind of network, in order to verify whether individual communities represent meaningful groups of personalities that might be used as an additional feature in text illustration or search in general. A similar network has recently been studied by Gargi et al. [3]. They focused on improving large-scale community detection using a multi-stage algorithm that combines a pre-processing stage, a local clustering stage, and a post-processing stage. Their YouTube graph, induced by the co-watching of videos, resembles the network we study here. In order to build the network, they considered that two videos which had a high co-watch value would be considered similar and therefore be connected by an edge. They imposed a threshold for the number of co-watched videos to keep the graph sparse. Given the considerably smaller size of our network and the stronger ties between people mentioned in a photo description, we considered that two personalities would be connected if they were coreferenced simply once.

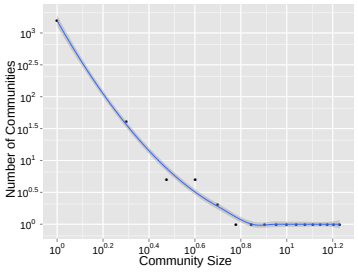
2 Building the Network

We use the SAPO Labs news stories photo collection [2]. This multimedia collection comprises over 1.5 million journalistic photos, with associated descriptions. Descriptions have an average length of 59 words, ranging from a minimum of 2

words to a maximum of 272 words. We parse the photo descriptions and, for each individual personality mention (according to SAPO Verbetes[1] GetPersonalities service), we store the photo ID and the personality name. Using R [4], we group by photo ID and build an adjacency list that includes duplicate personalities. We then eliminate duplicate edges, turning counts into weights, and obtain the adjacency matrix of the weighted, undirected personality coreference network, with 4,995 nodes and 15,929 edges. The network has a mean degree of 6.378, a median degree of 2, a diameter of 96 and an average path length of 3.89. 98.6% of the connected nodes are separated by a geodesic distance ranging between 2 and 6 edges. The network is sparse, having a density of 0.13%, and its clustering coefficient is 15.86%.



(a) Community structured.



(b) Community size distribution.

Fig. 1: The communities in the personality coreference network.

2.1 Community Structure

We identify the community structure of the network by using Gephi's implementation of the modularity optimization algorithm by Blondel et al. [1]. In Fig. 1a, we depict the different communities according to this algorithm. After obtaining the communities, we sampled some of the personalities of each community, discovered their occupation by searching Google and Wikipedia, and manually attributed a label to the communities based on the personalities professions. In order to validate the labels, we aggregated the news descriptions by

[1] `http://services.sapo.pt/Metadata/Service/InformationRetrieval/Verbetes`

community, removed Portuguese and English top words using Python's Whoosh *StandardAnalyzer*, and built the term frequency vector for each community's descriptions.

Fig. 1b depicts the community size distribution. We can see that it doesn't follow a power law, having a heavier tail behavior — the number of communities decreases more rapidly than an exponential function, as the community size increases. We extract the subnetwork for each community and analyze them individually. Table 1 illustrates some properties of the community subnetworks, including density, degree and PageRank. We verify that the Basketball community corresponds to the densest (12.12%) subnetwork. This means that it is the most interconnected community out of the seven communities. Notice however that this is still a low density value. The Soccer community has the largest mean and median degree, while the Basketball community has the largest mean and median PageRank score. This means that Soccer personalities are usually referenced alongside a large number of other Soccer personalities, while Basketball personalities are referenced alongside other important Basketball personalities, where important means that they are also referenced with other relevant Basketball personalities.

Table 2 depicts the top 5 most important personalities inside each community, according to degree and PageRank. Bold entities correspond to personalities that are out of topic, given the label assigned to the community. When doing a manual verification of the occupation of Axel Weber using Wikipedia, we found that Axel Weber could either be a retired East German pole vaulter or a German economist. Given he belongs to the Finance community, we could use this knowledge to disambiguate his occupation and conclude that this is actually a reference to Axel Weber, the German economist. We could say that, by identifying the community structure of this network, we obtain a context. For instance, in the Tennis/Formula 1 community, Thomas Gottschalk is in fact the host for a German entertainment television show called *Wetten, dass..?*. However, we could easily discover his connection to this community. By doing an online search, we found out that, in February 28th 2009, Boris Becker, a former professional tennis player from Germany, was a guest in Thomas Gottschalk's show. So, although empirically Thomas Gottschalk would be out of context in this community, we were able to find a simple explanation to justify the attribution. Additionally, this leads to the importance of considering the dynamics in communities, since the identified event that explains Gottschalk's community membership might as well be a rare event, contextually relevant only during a limited timespan.

3 Conclusions

We have analyzed a personality coreference network, identified its community structure based on modularity optimization, and studied the largest component of the network as well as its individual communities. The studied network shows all the typical characteristics of real networks, such as sparsity, scale freedom and small world phenomenon. We have observed that the network induced by the

Community Label	Density	Degree		PageRank	
		Mean	Median	Mean	Median
Portuguese Politics	0.60%	7.025	3.00	0.0008569	0.0004132
International Politics	1.22%	10.160	3.00	0.0011990	0.0005727
Finance	7.37%	4.866	3.00	0.014930	0.011330
Soccer	1.70%	**11.880**	**6.00**	0.0014310	0.0008550
Tennis / Formula 1	1.84%	6.195	3.00	0.0029590	0.0023430
Cycling	6.45%	3.614	2.00	0.017540	0.012920
Basketball	**12.12%**	4.000	3.00	**0.029410**	**0.021430**

Table 1: Community analysis (maximum values for each column are bold).

Rank	Personality	Degree	Personality	PageRank
1	Roger Federer	44	Roger Federer	0.014216810
2	Michael Schumacher	35	Michael Schumacher	0.012195036
3	Rafael Nadal	31	**Martin Scorsese**	0.010473509
4	Fernando Alonso	30	**Thomas Gottschalk**	0.010239316
5	Andy Roddick	27	**Cate Blanchett**	0.009569045

Table 2: Top 5 personalities for Tennis/Formula 1, ordered by Degree (left) and by PageRank (right).

coreference of personalities in news stories might be a good feature to include in search or text illustration in order to boost results, based on query references to personalities and the context provided by their communities. Although we have manually assessed the community membership for the highest degree and PageRank nodes in each community, as future work we propose that this should be done for the totality of the nodes, using automated methods based, for instance, on DBpedia's Ontology.

References

1. Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10008 (2008), http://iopscience.iop.org/1742-5468/2008/10/P10008
2. Coelho, F., Ribeiro, C.: Characterization of the SAPO-Labs News Stories Photo Collection (2011), http://www.inescporto.pt/~fcoelho/web/_media/files/2011sapolabs.pdf
3. Gargi, U., Lu, W., Mirrokni, V., Yoon, S.: Large-Scale Community Detection on YouTube for Topic Discovery and Exploration. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011). pp. 486–489 (2011), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2864/3256
4. R Development Core Team: R: A language and environment for statistical computing. In: R Foundation for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2011), http://www.r-project.org