Master DMKM Report

Mining parliamentary data and news articles to find patterns of collaboration between politicians and third party actors

Francisco Andrés RODRÍGUEZ DRUMOND

defended the July 7th, 2014

Supervision : Marta Arias, Universitat Politècnica de Catalunya,
Josep Larriba, Data Management group
(DAMA-UPC),

Location : Data Management group, Barcelona, Spain.

Abstract:

LOPEN IPSUM

Résumé :

LOPEN IPSUM

Contents

1 Hosting institution

DAMA-UPC, the DAta MAnagement group at Universitat Politècnica de Catalunya (UPC) is part of the Computer Architecture Department (DAC). The main research topics of DAMA-UPC are oriented to performance, exploration and quality in data management, focusing particularly on large data volumes. Specifically, they have investigated the creation of new data structures, algorithms, methods and applications in the area of Data Management that make it easier to manipulate large amounts of data.

DAMA-UPC is a member of Tecnio since 2005. Tecnio is an initiative of ACC10, the Agency for Innovation and Internationalization of the Catalan Enterprise, belonging to Generalitat de Catalunya. It is supported by Generalitat de Catalunya as a Consolidated Research Group (SGR-1187) and by the Ministry of Education and Science of Spain. Since its creation it has worked with important technological partners such as IBM, Oracle technologies, the Ministry of Cience and Innovation, the Health Department of the Generalitat de Catalunya, among others; and is current participating in four European Research Projects.

Moreover, DAMA-UPC has been a pioneer in scientific collaboration, creating and organizing for three consecutive years the Workshop on Graph-based Technologies and Applications (Graph-TA), organizing the 17th International Database Engineering & Applications Symposium (IDEAS 2013) and the First University-Industry Meeting on Graph Databases (UIM-GDB). The work of DAMA has also give birth to Sparksee, a graph database, and Sparsity Technologies.

[1]

[1] Text partially taken from `www.dama.upc.edu`

2 Acknowledgement

LOREM IPSUM

3 Introduction

Modern law-making is characterized by the indirect participation of third-parties in the legislative process to attempt to influence the contents and outcome of the discussions according to their interests. This process, known as *lobbying*, is regularly done by companies, Non-Governmental Organizations (NGOs) and in some cases citizens and foreign governments. For instance, when a law concerning human rights is being discussed in a parliament we can expect human rights NGOs to contact politicians to attempt to push their agenda.

Most of the times citizens are completely unaware of the lobbying activities of their representatives. This is due to the fact that it is hard to closely monitor the activities of politicians, to understand the intricacies behind the drafting of bills and to grasp the implications they have on organizations and society. Knowing how decisions makers are influenced is however of great interest for politologists, journalists and voters in general. By tracking these relations we would be able to understand better the way decisions are made by government, have a better sense of who politicians are, who they serve and what their agenda is, and in some cases fight corruption.

Recently, there has been considerable work in the development of data mining tools to analyze the way legislative bodies work. Social Network Analysis (SNA) has been widely used to understand the dynamics of complex social systems and is a natural tool for modelling the dynamics of power in a parliament. There is a wealth of knowledge to be extracted from parliamentary data to model relationships between politicians, identify key players and sub-communities, predict the voting of bills and in general gain a better understanding of the legislative process. However, finding political relations between representatives and third party actors, who often do not appear in bills or in the transcripts of debates, remains a challenge. How can we automatically detect that a politician and a company might be aligned from information available to the public?

News articles contain a sizable amount of information about what happens in a given country and about its relevant actors. Because of this, there is plenty of work by the scientific community to develop methods to automatically extract useful knowledge from news articles. The development of applications to automatically generate Social Networks (SN) from corpora of news articles is particularly interesting in our context. This has been done extensively in the past in different areas including defense, counter-terrorism, news summarization and crime prevention. The working assumption is that by detecting patterns of co-occurrence of two entities in text we can establish that they are in some way related.

In this study we use that assumption and formulate the hypothesis that by analyzing and combining parliamentary data and news articles we can generate graphs descriptive of the political closeness of politicians and relevant actors of a country. Our purpose is consequently to propose a method to automatically detect patterns that could be indicative of a lobbying activity. To do this, we automatically analyze news articles, text of bills discussed in a parliament and the transcripts of the debates. We show that it is possible to i) track the participation and the position of a politician with respect to a law and ii) find entities in news articles which could be directly related to the contents of a law. Laws thus can then be used as a cornerstone for detecting relationships between politicians and third-party actors. These relationships can then be used to build a SN which can be analysed using SNA tools. In this document we present our proposal and show the results obtained by its application in the context of the Parliament of Catalonia.

It is important to clarify that despite being motivated by the detection of patterns that can be suggestive of a lobbying process, in practice we aim to find relationships of political similarity or dissimilarity between two entities. The relations found by our method do not necessarily imply that the two actors are in direct liaison; to do so we would need to closely monitor all the activities of politicians and organizations to verify with whom they are in contact. This is naturally unreasonable due to privacy considerations.

We aim to detect links between entities that indicate that they are both related to a particular political decision - in our case bills -, from which we could then establish political affinity or aversion. Naturally, if two entities have highly similar political views in a broad set of issues then one can believe that they could be collaborating. This could be verified by investigative journalists and by considering other sources of information like donations information, speeches, etcetera. Regardless of the verification of a lobbying process, having a graph relating politicians and third parties is in itself useful to understand the political landscape of a country. In this study we also show how our proposal is useful for political analysis, besides from the interest in the detection of lobbying activities.

The rest of this document is organized as follows: in chapter 4 we present the state of the art in the automatic generation of SN from corpora of text and in the use of SNA in the context of political science. In chapter5 we present our proposal and in chapter 6 we present some implementation considerations deemed relevant. Next, we present the obtained results in section 7 along with some interesting applications of our method. Finally, we present in chapter 8 the conclusions of our work along with some suggestions for further work.

4 Related works

There is an increasing interest in the development of applications to generate Social Networks (SN) relating entities occurring in corpora of text. This is due to the fact that there is a growing wealth of knowledge contained in text documents which is difficult to exploit due to their unstructured nature. By generating graphs that subsume the information contained in these documents, we produce a structured view which can be analyzed using Social Network Analysis (SNA) tools.

In this section we present the state of the art in the generation of SN from corpora of text and some related applications that use SNA for political analysis. First in section 4.1 we present the entity co-occurrence approach, a simple technique that has been widely used with good results. We then show in 4.2 relevant work by the scientific community to characterize the link between two entities in a way that can be used by SNA. We also describe in 4.3 methods that can be used to relate entities that despite not being mentioned in the same document might be linked by means of the broader context provided by the whole corpus of documents. Next, we present in section 4.4 studies that use SNA for political analysis and that are related to our project. Finally, we succinctly present in 4.5 some considerations about how the state of the art was taken into account when making our proposal.

4.1 Entity co-occurrence: a first step towards the creation of SN

One of the most widely used approaches to generate SN from text consists in relating entities based on their co-occurrence in a certain context. The underlying assumption of this approach is that if two entities are consistently mentioned together then they are probably related. There are several variants depending on the granularity of the context; one can look for co-occurrence within a certain sentence, paragraph, document or cluster of documents. The choice depends on the amount of data available - finer granularity probably requires more documents to produce more relations - and on the application.

The authors of [4] propose a method to automatically generate a social network of narco-traffickers in Mexico based on the co-occurrence of names in books about the topic. They do Entity Recognition (ER) to produce a list of entities which is then manually curated and used to determine links between drug dealers. The weight of the relationship is the count of repetitions of the co-occurrences of two entities within a certain distance. They use different network analysis tools to show how the obtained graph closely resembles the different cartels and their chain of command.

Similarly, the Joint Research Centre of the European Commission has done extensive work in extracting entities and inter-entities relations from newspapers written in different countries of the European Union. In [11] we find a summary of their work, which is explained in depth in [9] and [12]. Essentially, they look for co-occurrence of entities within previously built clusters of articles that represent a story. They take into account entity coreference and use different heuristics to improve the entity recognition and disambiguation processes. They also produced a formula to measure the strength of a link based not only on the number of co-occurrences but also the frequency of the entities in the clusters and the corpus. By doing this, they aim to weight down relationships in which one of the entities is frequently mentioned, so that only relevant relationships are chosen.

Another interesting aspect of their proposal is the use of Wikipedia for validating the obtained graphs. Because we are in presence of a knowledge discovery task for which we do not have a ground truth set, it is difficult to evaluate if the detected links between two entities are meaningful. The definition of a meaningful link is itself not an easy task. The authors of the Joint Research Centre look for the Wikipedia pages of the detected entities and verify if there are links between pages that correspond to the links detected by the system. They define a "strong" relationship as one in which there is a reciprocal presence of a link.This allows for the creation of a ground truth set which can be used to evaluate the system with the standard precision and recall metrics.

On a different note, the authors of [2] present a technique to measure the semantic similarity of two words or phrases by using the Google search engine. They propose a metric based on information distance and Kolmogorov complexity that uses the count of search hits returned by looking up two words individually and together. They show how their approach is useful for distinguishing between colors, numbers, names of paintings and names of books, among others.

The main advantage of this approach is that it is able to measure the similarity of two entities based on the whole corpus of documents in the World Wide Web. The drawback is that the number of search hits returned by Google is a gross and often highly inaccurate estimate of the real count. Particularly, it is usually the case that a search with more terms returns a higher count of hits than a search with a subset of these terms. The reason for this is that when adding more terms the search is more fine-grained, allowing for a more refined estimate of documents. More specifically, when having more search terms it is necessary to go deeper through the posting lists which leads to more accurate and larger result estimates. The data centers or the indices used when answering the query also affect the number of expected hits returned. This makes approaches that depend too much on the exact count returned by the search engine unreliable.

There are two shortcomings in taking a co-occurrence approach. First, we are often interested in characterizing the link between two entities to produce richer graphs. It is true that the co-occurrence approach allows a human user to manually inspect the documents in which two entities co-occur. We are however particularly interested in mechanisms that can infer and represent the semantic nature of a link in a way exploitable by social network analysis tools with as little human participation as possible. Second, we are also interested in methods that do not rely on direct co-occurrence within a same document (or a pre-computed cluster of documents), but that can also discover meaningful relationships across a corpus.

4.2 What links two entities? Enriching the SN with the semantics of the relationships

As we previously said, the co-occurrence approach relies on the assumption that if two entities are mentioned in the same context then they are probably related. Co-occurrence may indeed be suggestive that two entities are related, but if there is a relationship we need to characterize that relationship before we can perform SNA. To illustrate this, in the context of our application finding a link between a politician and a third party might indicate that they are closely aligned politically, that they are in opposition, that they participated together in a meeting, that they mentioned each other, among others. Having more information about the found relationships is consequently of great importance to be able to do better analysis.

The efforts of the scientific community to characterize relationships between two entities have been mostly concentrated on Natural Language Processing methods to analyze the documents. The authors of [13] propose a method which uses dependency trees to learn patterns that relate two entities co-occurring in a sentence according to a pre-defined type of relationship. They work with two examples of relationships: " support" and "meeting". By working with a small, manually obtained number of seed instances of the relationship - tuples of entities -, they look for sentence co-occurrence in a group of news articles and extract patterns by using their SyntNet GSL algorithm over the dependencies found by a dependency parser.

Similarly, in [10] the authors propose a method to automatically detect quotation relationships in news articles. They aim to do this by also finding linguistic patterns that are usually used when expressing citations: quotation markers and reporting verbs.

The authors of [15] illustrate the use of kernel methods for relation extraction. They first produce a shallow parse representation of the texts which is then used by kernels designed specifically to work on

parse trees, which have been defined in [3]. These kernels are able to implicitly enumerate all possible subtrees of two parse trees, find which are the most common subtrees, weight them and compute a similarity measure based on these. By using a pre-obtained ground truth set, they are able to train classifiers to determine, given two entities and a tree describing the sentence they co-occur in, if the entities have relationships of the type person-affiliation and organization-location.

4.3 Going beyond the co-occurrence approach: finding links between entities across documents

There is also a number of techniques to address the need to find links between entities without depending on their direct co-occurrence within a same document or pre-computed cluster of documents. A widely used approach is automatically inferring topics present in a corpus of text and to verify co-occurrence in articles related to these topics. The authors of [8] propose a method that uses Latent Dirichlet Allocation (LDA) to produce a topic model of the documents. In LDA a document is regarded as a finite mixture of topics and represented as a vector in which each component constitutes the probability that the document belongs to a given topic. This model is then used to calculate an entity-entity measure of affinity that is used to find links between entities. The reported results are motivating; the authors show how the use of topics allows to discover more links between entities and to characterize them by means of the topics. They also report however that LDA has however one significant disadvantage: the obtained topics may be hard to interpret and may not be sufficiently semantically cohesive.

An alternative to the use of LDA is Latent Semantic Indexing (LSI). By producing a vectorized representation of entities (in which for instance we store information about their co-occurrence in documents), we can use Singular Value Decomposition (SVD) to find a lower dimensional space and estimate the similarity of entities based on latent concepts. This is useful for noise-reduction and for finding semantic relationships between entities. The drawback is that the found relationship may be hard to interpret. The authors of [1] provide an example of the use of LSI for the generation of graphs of terrorists networks.

4.4 SNA and political analysis. Has anyone done this before?

To the best of our knowledge, there are no approaches in the literature for finding patterns that could be indicative of a lobbying process in an automatic fashion. There are however several studies that use SNA for studying the dynamics of power, particularly in legislative bodies. For example, the authors of [5] studied a graph of co-sponsorship of bills to study interactions between congressmen in the US House of Representatives and found that by using network analysis tools it was possible to find highly influential politicians. Similarly, in [6] the authors developed a theory of influence diffusion across a legislative network of relations based on weak and strong links and found patterns useful for determining the success of a bill. In [14] we find a proposal to predict the voting of a bill based on speeches made by congressmen.

In [7] we find a proposal for the automatic generation of policy networks that is highly related to our work. They work with two SNs relating political actors previously created by experts in a manual, time-consuming process. They evaluate the use of co-occurrence measures in different types of contexts (page-counts vs co-occurrence within a certain distance) and the use of link metrics (hyperlinks between web pages) by using the Yahoo search engine to generate graphs and verify their overlapping with the manually generated SNs. The manually created SNs contain positive and negative edges, which correspond to relationships of political affinity and aversion, and are useful for understanding how the different methods for graph generations perform.

In general, they obtained better results for when detecting affinity relationships than negative relationships (a correlation of up to 0.74 is achieved). They also find that using link metrics is the best

alternative for positive relationship while context-based metrics are the best option for negative relationships. Among the main difficulties they found they distinguish data sparseness, actor name ambiguity, language, and relation type. The main difference with our work is that while they work with a predefined list of political actors from the policy networks they use for validation, we are also interested in the discovery of relevant entities and in finding ways to characterize their links. The authors of this proposal did not present any alternatives for automatically inferring the sign of the relationships detected by their system.

4.5 Framing our work with respect to the state of the art

As we have seen, generating social networks from corpora of text is a topic widely addressed by the scientific community. There are two important considerations with which researchers are concerned: i) discovering the largest number of relationships possible while ensuring the discovered relationships are meaningful and ii) characterizing the relationships between entities in a way exploitable by SNA tools.

In the context of our application we are interested in discovering meaningful relationships between politicians and third-parties. We define a meaningful relationship between these two types of entities as a relation of political closeness, meaning that they are affected and have established positions on laws, decrees and other political decisions taken by a parliament. We present in depth our proposal in section [**?**]. However it is worthwhile to present in this section some considerations concerning the state of the art that were taken into account when making our proposal.

The task of discovering patterns of political closeness between politicians and third parties is conditioned by the fact that these relationships are usually hidden and unknown by journalists and the public. Entity co-occurrence in a given context thus entails a risk of not revealing all the relevant relationships. Similarly, characterizing the relationships by means of NLP techniques would also require knowledge by the writer of the document of the existence of a relationship between two entities. Finally, discovering topics from the corpus of documents could also lead to using topics that do not correspond to political themes.

To address this, we use bills as the cornerstone that allows us to link politicians and third parties. By using bills, we can model interpretable and semantically cohesive topics that allows us to address the two considerations mentioned earlier in this section. We can specifically i) detect meaningful links between entities across documents - thus increasing the precision and recall of our system in terms of an imaginary ground truth set - and ii) characterize the found links by using the bill that led to the discovery of the link.

5 Proposal

Here I explain the proposal without implementation considerations.

6 Implementation

In this chapter give some relevant implementation details, along with a diagram showing the pipeline of the system.

6.1 Modelling topics after bills

6.2 Entity Recognition and Preprocessing

After finding articles related to bills, they are analyzed to find entities that are in turn related to the bills. For this purpose we use MITIE, a state of art tool Named Entity Recognition tool created in the MIT. Given a document, MITIE identifies substrings that contain possible named entities and tags them as *Organization, Location, Person* or *Miscellaneous.*

Before carrying on the detected entities need to be pre-processed before they can be used. The names found may i) not be correctly delimited (resulting in truncated names or in names that contain excess text) ii) be ambiguous (an entity may have more than one name and a name may refer to more than one entity) and iii) may be noise and not refer to a real entity.

Name disambiguation is in itself a complex and interesting research topic which escapes the aim of this study. We have however implemented some heuristics we briefly describe below:

1. **Entity Normalization**: entities are brought to a canonical form to address spelling variations. This involves i) punctuation sign removal; ii) double, leading and trailing whitespaces removal; iii) leading and trailing stop words removal; iv) string camelization (all characters are put in lowercase except for the first letter of every word, which is in uppercase).

2. **Mapping organization initials to the whole name**: when an organization name is detected we aim to detect if there are any other names composed by its initials. Specifically, we aim to exploit a widely found pattern: organizations often have the full name followed by the initials inside parenthesis.

3. **Mapping partial names with full names**: when person names are detected we classify them into *full names* (containing more than 1 word) and *short names* (containing one word, which is possible the last or first name of the full name). We then link short names with the nearest, previous full name such that the short name is contained inside the full name.

4. **Expanding names based on the news corpus**: to address the issue of truncated names (eg 'Word Life Fund' may be truncated and processed as 'World and 'Life Fund') we: i) look up every name and it's surrounding context in the corpus ii) extract sentences in the top articles and iii) find the longest substring matching these sentences.

By doing this we find a list of entities for which we have a list of aliases and a list of tags counts. After doing this, we choose for every entity the tag with the highest frequency as the type of the entity.

7 Results

Results go here.

8 Conclusions

Conclusions come here.

References

[1] Roger B Bradford. Application of latent semantic indexing in generating graphs of terrorist networks. In *Intelligence and Security Informatics*, pages 674–675. Springer, 2006.

[2] Rudi L Cilibrasi and Paul MB Vitanyi. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383, 2007.

[3] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In *Advances in neural information processing systems*, pages 625–632, 2001.

[4] Jesús Espinal-Enríquez, J Mario Siqueiros-García, Rodrigo García-Herrera, and Sergio Antonio Alcalá-Corona. A literature-based approach to a narco-network. In *Social Informatics*, pages 97–101. Springer, 2014.

[5] James H Fowler. Connecting the congress: A study of cosponsorship networks. *Political Analysis*, 14(4):456–487, 2006.

[6] Justin H Kirkland. The relational determinants of legislative outcomes: Strong and weak ties between legislators. *The Journal of Politics*, 73(03):887–898, 2011.

[7] Theodosis Moschopoulos, Elias Iosif, Leeda Demetropoulou, Alexandros Potamianos, and Shrikanth Shri Narayanan. Toward the automatic extraction of policy networks using web links and documents. *Knowledge and Data Engineering, IEEE Transactions on*, 25(10):2404–2417, 2013.

[8] David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Analyzing entities and topics in news articles using statistical topic models. In *Intelligence and Security Informatics*, pages 93–104. Springer, 2006.

[9] Bruno Pouliquen, Ralf Steinberger, and Jenya Belyaeva. Multilingual multi-document continuously-updated social networks. In *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pages 25–32, 2007.

[10] Bruno Pouliquen, Ralf Steinberger, and Clive Best. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*, pages 487–492, 2007.

[11] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tamara Oellinger. *arXiv preprint cs/0609066*, 2006.

[12] Bruno Pouliquen, Hristo Tanev, and Martin Atkinson. Extracting and learning social networks out of multilingual news. In *Proceedings of the Social Networks and Application tools workshop (Skalica, Slovakia, Septembe*. Citeseer, 2008.

[13] Hristo Tanev. Unsupervised learning of social networks from a multiple-source news corpus. *MuLTISOuRcE, MuLTILINguAL INfORMATION ExTRAc-TION ANd SuMMARIzATION*, page 33, 2007.

[14] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics, 2006.

[15] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106, 2003.

ANNEXES