

Extracting and Learning Social Networks out of Multilingual News

Bruno Pouliquen¹, Hristo Tanev¹, Martin Atkinson¹

¹ Joint Research Centre - European Commission

Via E. Fermi 2749, 21027 Ispra, Italy

{Bruno.Pouliquen, Hristo.Tanev, Martin.Atkinson}@jrc.it

Abstract. Various kinds of social networks can be derived from the analysis of news articles. We present here our experience in building social networks by the extraction of relationships between entities all automatically derived from multilingual news articles. Unqualified relationships between persons can be extracted through simple co-occurrence statistics. Qualified relationships can be extracted using linguistic patterns. Our highly redundant sources (50,000 daily articles in 40 languages) are used to both validate our algorithms and strengthen pertinent relationships. Due to the amount of data we process these social networks provide a complex challenge for their useful visualization and navigation.

Keywords: Link Analysis, Knowledge Acquisition

1 Introduction

Our system Europe Media Monitor (EMM) gathers an average of 50.000 multilingual newspaper articles daily. To manage the problem of information overflow we provide users with fully automatic tools that summarize long-term and live information in multilingual news.

One of the user requests concerns deriving social networks as they appear in the news. Manual building of such social networks is unfeasible on a wide scale. There is a clear need for an automatic approach to extract Social Networks out of unstructured information such as multilingual news.

EMM collects information in currently 40 languages from about 1.500 websites around the world. This information is highly multilingual and highly redundant (a lot of events may be reported in different sources and different languages). While tackling so many sources and languages is quite challenging it is also a clear advantage because it provides statistical evidence for the relative importance of the relationship between two entities.

Most of the information available in newspapers mentions subjects involving entities (mainly persons, organizations and places). We have a tool that extracts entities automatically from unstructured texts (see Section 2.3). Each incoming article is tagged with the entities it mentions. Several techniques can then be applied to infer relationships between entities. Two main families of relationships can be inferred: unqualified co-occurrence relationships (simply based on the fact that two entities appear in the same article or group of articles) and qualified relationships (based on some patterns that try to “understand” the sentence where the two entities appear).

Unqualified co-occurrence relationship relies on statistics that are easy to extract in our environment because it does not require language-dependent knowledge. The qualified relationship relies on a deeper analysis of the text to extract recognizable patterns to infer a category of relationship (meeting, quote, family relation etc.).

When two entities appear in the same article we can then create a link which we compile in “live social networks” (see section 3.1). Similar articles are grouped into clusters every day. Entities appearing often in the same clusters create “long-term relationships” (see section 3.2).

A deeper analysis of articles can also reveal links between these persons. We construct syntax-based Social Networks between persons (section 4.1). Similarly we construct social networks out of reported persons in quotations (section 4.2).

When entity relationships are extracted we try to summarize the information in graphs in order to reveal social interactions between entities. Building and displaying such graphs is a strong user requirement, section 5 will describe our experience regarding this issue.

2 Background: related work, data and tools

2.1 Related work

Social networks are sometimes derived from statistical data like Matsuo et al. 2007, where the advantage is the amount of data that can be processed and the disadvantage is that they cannot qualify the type of extracted relation. Other test-based relation extraction tools like Zelenko et al. (2007) or Romano et al. (2007) give more precise relationships (a single sentence can create a relationship). The cost is the amount of linguistic resources required and a problem of recall due to the fact that an expression will not be recognized if it is not pre-defined in some rule.

Ben-Dov et al. 2004 compared the two methods: statistic and rule-based. They came to the conclusion that statistics-based methods are good at recall and bad at precision, while rule-based methods are good at precision and bad at recall. They estimated that a rule-based algorithm required from one to three weeks to write the rules for detecting a ‘participation in a common meeting’ (with an existing syntactic analysis tool).

Other examples of research in the wide field of text link analysis can be seen at the Workshop on Text-Mining & Link-Analysis (TextLink 2007) held at the IJCAI conference. More specifically examples about link analysis can be found in the Workshop proceedings on Link Analysis: Dynamics and Static of Large Networks (LinkKDD2006)¹.

¹ <http://kt.ijs.si/dunja/textlink2007/>, <http://kt.ijs.si/Dunja/LinkKDD2006/>

2.2 Europe Media Monitor

The EMM project was started six years ago and currently gathers 50.000 newspapers articles in 40 languages from more than 1.500 websites around the world. A process groups articles talking about the same subject into “clusters”. This data can be browsed through the NewsBrief page for live news (<http://press.jrc.it>) or through the NewsExplorer page for long term analysis of news (<http://press.jrc.it/NewsExplorer>).

2.3 Named Entity Recognition

We built our own system for Named Entity Recognition. This system is fully described in Steinberger & Pouliquen (2007). To summarise, we guess new person names using lists of first names and trigger words (like “Mr.”, “Minister”, “playboy” etc.). This is currently done in 19 languages (including Arabic, Russian, Bulgarian etc.). The new names (665 per day on average) are stored in a knowledge base together with their context (e.g. which date and which cluster they appeared in). It is quite important to highlight the fact that we do recognise multiple variants of the same name (even across scripts). On average 91 out of 665 names (14%) are variants of a person already in our knowledge base. For the Libyan leader Muammar al-Gaddafi we collected more than 100 variants².

Once a day, we compile lists of persons and organisations as finite state automata to look up names in every single EMM article or cluster of articles.

3 Statistics-based relationship

3.1 Live social networks

This work is fully described in Pouliquen et al. (2007a). The aim is to build social networks of persons mentioned in the news of the last few hours. This fully automatic system is online at the following address: <http://langtech.jrc.it/SocNet/>

Each incoming article in EMM is tagged with the known named entities it contains. We create a relationship between each entity pair by summing the number of articles where they appear together. For efficiency reason this index is kept simple as it is updated every second and read every two hours (we reset it at midnight every day).

This index of relationships is then turned into a network where each node (vertex) is an entity and each link (edge) is weighted by the number of articles where the two entities appear together. In addition we provide a hyperlink to the context for each single link (i.e. to the original article and a snippet showing the text snippet where each name appears, see Figure 1).

² <http://press.jrc.it/NewsExplorer/entities/en/262.html>

2007-07-13T06:38+0200 . [UN nuke delegation arrives in Iran](#) [iranmania] ...to meet with Iran's top nuclear negotiator, [Ali Larijani](#), later in the day, the report said. Accordiated Press (AP), Larijani and IAEA Chief [Mohammad ElBaradei](#) met last month in Vienna, Austria. Earli...
 LONDON, July 12 (IranMania) - Iran's President Mahmoud Ahmadinejad said that the West should not expect his country to suspend uranium enrichment activities, the official Islamic Republic News Agency reported.

2007-07-13T06:37+0200 . [OIEA asegura haber logrado acuerdos Irán](#) [HoyDigital-DO] ...Internacional de Energía Atómica (OIEA), [Mohamad el Baradei](#), hizo esta declaración al término de la (...) i, el asesor del principal negociador iraní [Ali Larijani](#), que preside la parte iraní en las negociac...
 TEHRAN, (EFE).- El jefe de la delegación del OIEA que visita Irán, Olli Heinonen, afirmó ayer que su equipo ha alcanzado un acuerdo con los dirigentes iraníes sobre "algunas cuestiones" en las negociaciones entre las dos partes sobre el caso nuclear iraní.

2007-07-13T02:13+0200 . [R E G I O N : Iran, UN team hold talks on nuclear issues](#) [dailytimesPK] ... deputy to Iran's chief nuclear negotiator, [Ali Larijani](#). President Mahmoud Ahmadinejad said on Wedn (...) unciil. The UN watchdog's Director General [Mohamed ElBaradei](#) has said Iran's transparency offer combi...
 TEHRAN: Iranian nuclear officials and a visiting team from the UN nuclear watchdog held a second round of talks on Thursday to discuss ways to remove outstanding questions about Iran's disputed nuclear programme. Iran has offered to draw up an "action plan" to address Western suspicions that its nuclear programme is a front to obtain nuclear arms.

2007-07-13T01:31+0200 . [نتائج بداية بين إيران والوكالة الدولية للطاقة](#) [alrai] ... في حين أعلن مسؤولون إيرانيون في طهران أنهم...
 طهران - وكالات - أعلنت إيران أمس انه تم التوصل الى نتائج جيدة بعد ثلاث جولات من المحادثات مع وفد من الوكالة الدولية للطاقة الذرية. وابتعدت الجولة الثالثة من المحادثات حول البرنامج النووي الإيراني بين المسؤولين الإيرانيين ووفد الوكالة الدولية للطاقة الذرية برئاسة لوي عابونين نائب مدير الوكالة الدولية للطاقة الذرية.

Figure 1: Co-occurrence context between two entities (M. ElBaradei / A. Larijani)

The resulting graph is usually too big to be displayed as such to users. We filter, split and reduce it so that the user can easily browse it in an intuitive way as shown in Figure 2.



Figure 2: “Live” social network example

3.2 Building long-term relationships between persons

Every day the articles are grouped by language into “clusters” (groups of similar articles). These clusters and the persons they mention are stored in a knowledge base which allows us to extract fine-grained information such as “all persons appearing often in clusters together with a given person in a certain time period”.

The result is available for the user on the NewsExplorer website (available at: <http://press.jrc.it/NewsExplorer/>). Each person has its own page showing the relations he/she has with other persons or organizations.

Similar to the previous section (live social networks), we compute the relationship between names using a simple co-occurrence frequency at the cluster level. This

relation is called “related” people. The “related” person links are then compiled in a social network. It must be highlighted that the social networks are displayed currently using a window of 365 days of news, which means that some people may disappear if they are not mentioned in the news any more. Inversely, when new people appear in the news they are added to the network.

Some people are always in the news and may appear in a lot of “related” links. We therefore introduced a new formula to compute the “associated” people. This relationship gives higher weight to people that specifically appear together.

These two relationships are displayed on NewsExplorer entity pages.

Combining various relationships creates social networks. As we currently have about 150.000 persons having relationships it would be difficult to display all relations in a single graph. We currently display social networks only on a per-person graph, as shown in the two examples in Figure 3.

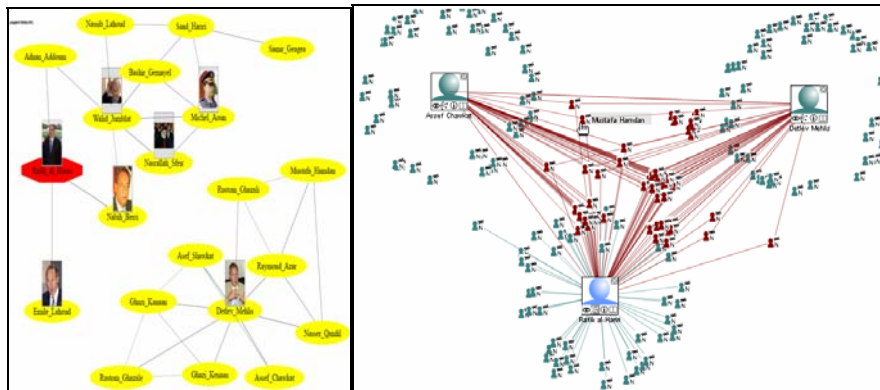


Figure 3: Social network examples

4 Linguistic patterns-based relationship

4.1 Syntax-based relationship (in political analyst tasks)

A complex syntax-based Social Network learning algorithm was described earlier in Tanev, 2007. It begins with a couple of syntactic patterns such as “*PERSON1* meets *PERSON2*” and then learns syntactic paraphrases, such as “*PERSON1* has a dinner with *PERSON2*” (here we show linear forms of the paraphrases, however they are syntactic graphs). To improve the accuracy of the pattern library, we performed a manual check-up. Next, with the learned syntactic patterns about meetings, contacts,

expressed support, criticism and family relations, our system extracts social relations from the EMM news.

We tried to use the automatically extracted social network of contacts via two applications intended to facilitate some political analysis tasks.

The first one was reported earlier in Tanev 2007. It uses the Eigen-vector centrality measure (used also in the PageRank algorithm) to estimate the importance of a political person. In Table 1 it is shown the top 5 politicians according to two ranking schemas - Eigen-vector centrality and frequency of mentions. The corpus from which we extracted the data covered one month – October, 2006. The Eigen-vector centrality ranks *Condoleezza Rice* higher than *George W. Bush* and *Tony Blair*, since she was on an important tour in the Middle East in the period covered by the test corpus. The most important difference in the two ranking schemas is that the Eigen vector centrality ranks in the top 5 *Vladimir Putin* and *Ehud Olmert* – two very important political leaders, while frequency ranking fails to do this and gives preference to *Saddam Hussein*, which during the considered period was in the jail and had no importance for the World politics.

Rank	Eigen vector centrality ranking	Frequency based ranking
1	Condoleezza Rice	George Bush
2	George W. Bush	Tony Blair
3	Vladimir Putin	Condoleezza Rice
4	Ehud Olmert	Nouri al-Maliki
5	Tony Blair	Saddam Hussein

Table 1: top 5 ranking politicians

The second application follows the development in time of the media reports about the meetings and contacts of a political person. We experimented with this application for several well-known politicians, namely *Fouad Siniora*, *Kofi Annan*, *Hugo Chavez*, *Condoleezza Rice*, *Hosni Mubarak* and *Olsegun Obasanjo*.

For each person we extracted from our database all the automatically detected media-reported contacts (mostly meetings, but also phone conversations and exchanged letters) in the period 01 January 2005 – 01 August 2007. This period was divided into sub-periods of three months and for each such sub-period we calculated the number of the different people with whom the political person of interest had a meeting or another type of contact. For each of the politicians, mentioned above, we presented the number of contacts in a graph which follows their development over time. See, for example the graph for the former President of Lebanon *Fouad Siniora* in Figure 4.

It turned out that for most of the politicians the peaks and the minima in their corresponding contact graphs coincide with important political events in which these people have an outstanding role. For example, the maximum in the contacts of the Lebanon president in Figure 4 is in the post-war period after the 2006 Lebanon-Israel war, when the peace between Israel and Lebanon was being established.

However, it turned out that curves which reflect frequency of mentions during the same periods have similar minima and maxima for most of the aforementioned political leaders. Nevertheless, there were some important differences. For example, the contact graph of *Condoleezza Rice* has a clearly distinguished maximum for

August-October 2006, which coincides with the aforementioned *Rice's* Middle East tour. The frequency graph does not have such a maximum. Another example is the re-election of *Hosni Mubarak* in September 2005, which produces a maximum in the contact graph of this politician, but fails to be reflected in the corresponding frequency graph.

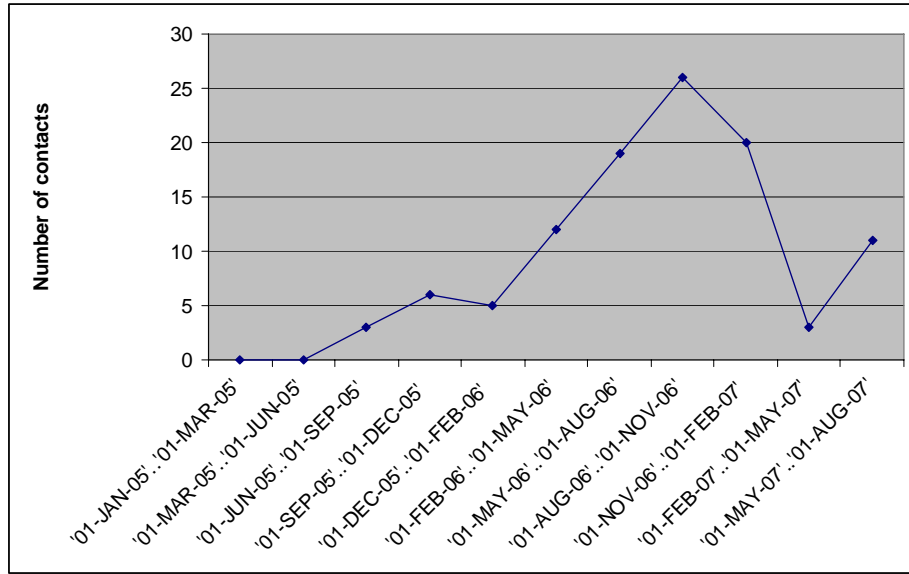


Figure 4: **Development of the contacts of *Fouad Siniora***

These observations make us think that the contact graphs can be used to complement the information from the frequency graphs to detect time periods important for certain politicians.

Our experiments show that automatically extracted social network of contacts may be used for finding the importance of political persons and their evolution over time. In some cases the information derived from the social network may be considered more reliable than information derived from frequency analysis. Social network and frequency analyses may also be combined to facilitate some political analyst tasks.

4.2 Quotation relationships

This particular relationship (a person mentions another person in his/her reported speech) relies on light language resources and is currently running in 16 languages. The system is fully described in Pouliquen et al. (2007b).

Relying mainly on a list of reporting verb (said, declared, etc.) and quotation markers (“”, «», etc.) the system takes advantage of recognized entities in texts to extract direct speech quotations. For example, from an article containing:

Betancourt's daughter Melanie Delloye stood next to French President Nicolas Sarkozy in Paris. "Today there is immense joy. All of France is happy about the rescue of Ingrid Betancourt," said Sarkozy.

The algorithm extracts the quotation, assigns it to the person *Nicolas Sarkozy* and records the fact that the quotation contains another person, *Ingrid Betancourt*. Thus we can create a directed relationship between *Sarkozy* and *Betancourt* qualified as “quoted”.

Every day a process extracts from the knowledge base all “quoted” relationship between persons (in a one-month period of time) and builds a social network where every node is qualified by the quotation and a link to the original article containing the quotation. We presently keep only one quotation. When more than one are available we cluster them and select the medoid quotation. The result is available online at <http://langtech.jrc.it/entities/socNet/quotes.html>

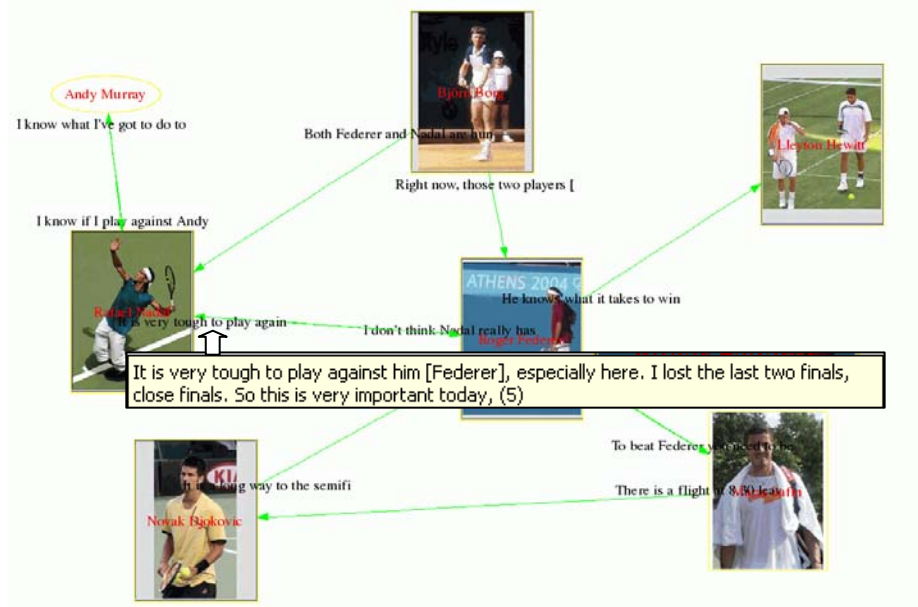


Figure 5: Quotation network example

5 Visualisation and Navigation through Social Networks

The main challenge in visualizing these generated social networks is that there are often a large number of nodes and links. Some of our graphs have more than 7,000 nodes and 13,000 links. Clearly it's not sensible or feasible to show the complete

graph on a single screen. Instead we offer the possibility of visualizing a sub-graph and then the option to explore the graph by navigating along visualized links.

In the past we used *Flash* interactive maps and *graphviz* static maps (see examples in Figure 3) as described in Pouliquen et al. (2006). These tools have limited functionality. We are now moving to a more sophisticated approach.

In our experimental system we use a Rich Internet Application client based on the Adobe Flex library. In particular we use a component called Spring Graph developed by Mark Shepherd from Adobe Labs³. This component forms the basic engine for the graph visualization. We also use a "type ahead" component that allows a user to select the starting point of the subgraph. This shows the names of all the nodes that match the typed string after every keystroke. Once a starting node is selected, a subgraph is visualized. The subgraph consists of all the connected nodes plus their connections for a given number of hops or steps across links. The number of hops is configurable since the ideal number depends on the density or connectivity of the node relations.

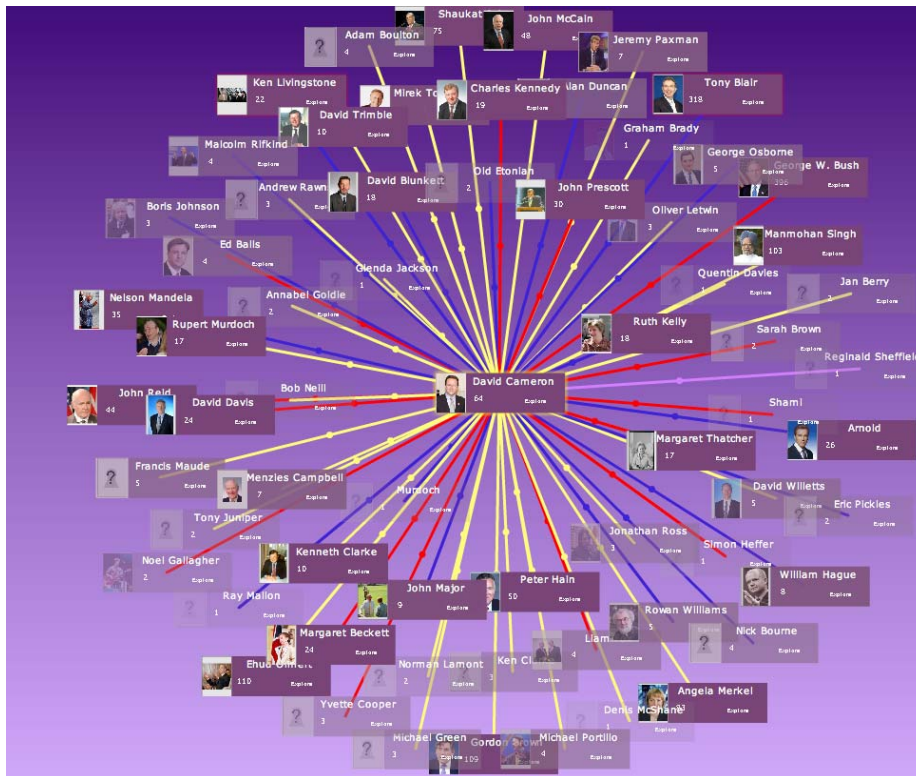


Figure 6: Links around British politician *David Cameron*

³ Component and documentation available at <http://mark-shepherd.com/blog/springgraph-flex-component/> (last visited 12/08/2008)

Navigation is achieved through the selection of the next node to focus on either through the "type ahead" component or by selecting a visualized node. Once a new node is selected it becomes the centre of the view and only nodes to the given number of hops are visualized.

A common problem is that some nodes have many links (these tend to be politicians - like *David Cameron* – see figure 6). Here to permit a faster or clearer visualization we allow filtering based on connection type. It is then possible to eliminate from the view all or show only the nodes that have a certain type of relation.

Finally in order to intuitively visualize the importance of relations we use the weighting of the relationship link in the automatic layout algorithm to draw stronger relations closer together and weaker relations further apart.

Figure 7 shows an example of the graph with 1 hop selected and the node *David Cameron*. On one side is the graph with all links selected and the left shows the criticize relations with hops set to 3

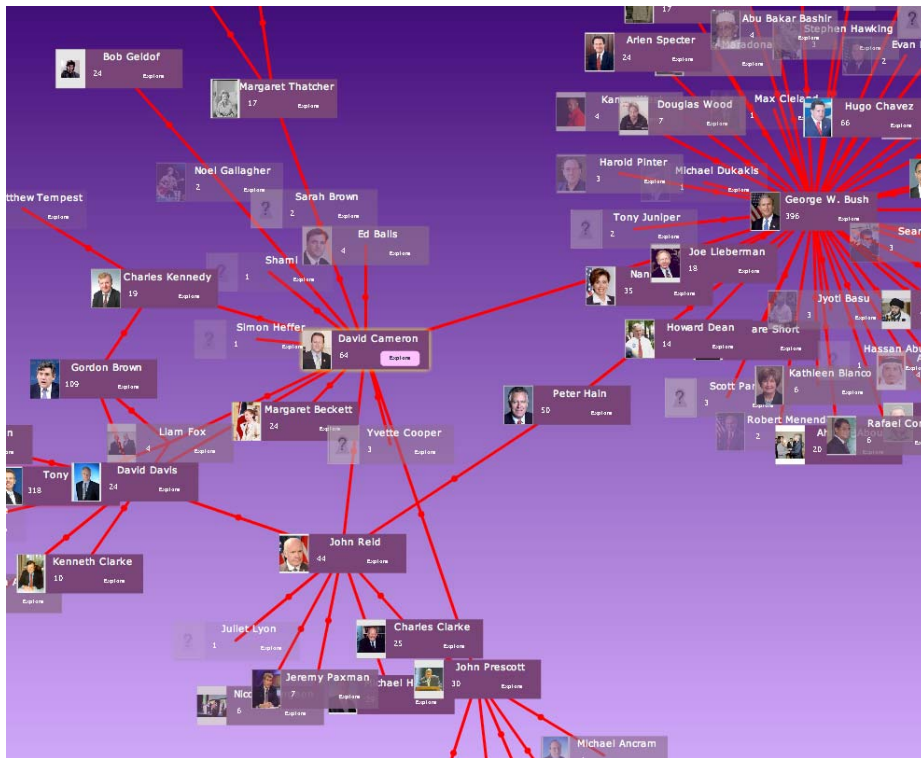


Figure 7: Visualisation example: *David Cameron*'s criticize social network

This visualisation tool is available at <http://emm-labs.jrc.it/LiveNews/Nets/Social.html>

6 Conclusion

Various approaches can be applied to extract social networks from multilingual news. Social networks building and visualisation can be used to derive compact information from multiple sources and even multiple languages. Our website <http://press.jrc.it/> is currently very popular and social network browsing is an important component (2000 hits per day for our current social network browser).

It goes without saying that we do not provide ready-made socio political analysis systems. These systems should be seen as tools that may help analysts do their work.

Acknowledgements

Many thanks to the whole team of the Web mining and Intelligence Action, especially to Ralf Steinberger, Jenya Belyaeva and Erik Van der Goot.

References

1. Tanev, H.: Unsupervised Learning of Social Networks from a Multiple-Source News Corpus. Proceedings of the Workshop *Multi-source Multilingual Information Extraction and Summarization* (MMIES'2007) held at RANLP'2007, pp. 33-40. Borovets, Bulgaria (2007)
2. Poulighen, B., Steinberger, R., Belyaeva, J.: Multilingual multi-document continuously updated social networks. Proceedings of the Workshop *Multi-source Multilingual Information Extraction and Summarization* (MMIES'2007) held at RANLP'2007, pp. 25-32. Borovets, Bulgaria (2007a)
3. Poulighen, B., Steinberger, R., Ignat, C. & Oellinger, T.. Building and displaying name relations using automatic unsupervised analysis of newspaper articles. Proceedings of the 8th International Conference on the Statistical Analysis of Textual Data (JADT'2006). Besançon, (2006)
4. Oezden Wennerberg, Pinar (2007). Analyzing Social Networks in Online News Articles. In: Norbert Gronau & Claudia Müller (eds.): *Analyse sozialer Netzwerke und Social Software -Grundlagen und Anwendungsbeispiele*, pp. 157-184. GITO-Verlag - Expertenwissen für die industrielle Praxis, Berlin
5. Steinberger, R., Poulighen, B., Cross-lingual Named Entity Recognition. In: Satoshi Sekine & Elisabete Ranchhod (eds.), *Journal Linguisticae Investigationes*, Special Issue on Named Entity Recognition and Categorisation, LI 30:1, pp. 135-162. John Benjamins Publishing Company. ISSN 0378-4169. (2007)
6. Poulighen, B., Steinberger, R., Best, C.: Automatic Detection of Quotations in Multilingual News. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'2007), pp. 487-492. Borovets, Bulgaria (2007b)
7. Matsuo, Y., Mori, J., Hamasaki, M., and Ishida, K.: POLYPHONET: An Advanced Social Network Extraction System from the Web, Proceedings of WWW conference (2006)
8. Zelenko, D., Aone, C., Richardella, A.: Kernel Methods for Relation Extraction, *Journal of Machine Learning Research*, vol.3, (2007)
9. Romano, L., Kouylekov, M., Szpektor, I., Dagan, I., and Lavelli, A.: Investigating a Generic Paraphrase-based approach for Relation Extraction, EACL, Trento, Italy, (2006)
10. M. Ben-Dov, W. Wu, R. Feldman & P. Cairns, Improving knowledge discovery by combining text-mining and link analysis techniques. SIAM International. Conference. on Data Mining, Florida, USA (2004)