A system to extract social networks based on the processing of information obtained from Internet

Xavi CANALETA\textsuperscript{a}, Pablo ROS\textsuperscript{b}, Alex VALLEJO\textsuperscript{b}, David VERNET\textsuperscript{c}, and Agustín ZABALLOS\textsuperscript{b}

\textsuperscript{a}Grup de Recerca en Sistemes Distribuïts
\textsuperscript{b}Secció de Telemàtica
\textsuperscript{c}Grup de Recerca en Sistemes Intel·ligents
Enginyeria i Arquitectura La Salle – Universitat Ramon Llull
\{xavic,pros,avallejo,dave,zaballos\}@salle.url.edu

Abstract. The amount of information available on Internet increases by the day. This constant growth makes it difficult to distinguish between information which is relevant and which is not, particularly in key areas which are of strategic interest to business and universities or research centres. This article presents an automatic system to extract social networks: a software designed to generate social networks by exploiting information which is already available on Internet through the use of common search engines such as Google or Yahoo. Furthermore, the metrics used to calculate the affinities between different actors in order to confirm the real ties which connect them and the methods to extract the semantic descriptions of each actor are presented. This represents the innovative part of this research.

Keywords. Social networks, web mining, web crawling, affinity.

Introduction

A social network is defined as a specific group of actors who are linked to each other through a series of social ties. The strength of these ties may vary depending on the actors in question.

A social network can be configured by obtaining information from each of the individual actors who might be in some way connected. Surveys carried out on different individuals, a behavioural analysis of the community and a data collection of each candidate provide a topology of the social network and show the strength of relation between the actors.

A clear example of the configuration of social networks is the creation of virtual communities by means of Internet portals. The appearance of Web 2.0 has placed a much greater emphasis on these user groups. The origin of this type of networks dates back to 1995 when Randy Conrads created Classmates. The year 2002 saw the beginning of web sites which promoted the creation of communities of friends online. And, as from 2003, the emergence of Friendster, MySpace or LinkedIn has made these types of sites popular. Currently, one of the most popular social networks in Spain is
Facebook. But the extracting system of these classic networks is not automatic. That is, the information required from each actor is provided by the individual himself.

The purpose of our line of work is that the extracting system of the social network is automatic. The information required from each actor is provided by the system which is fed in an automated form by Internet to compile previously stored data. The main difficulty lies in ensuring that the information extracted from each actor is of good enough quality. To achieve this, we have to address the problems related to the immense quantity of information which is difficult to classify [1], to the difficulty of identifying the actors [2], and the use of appropriate formulae to connect the actors of the social network to establish corresponding degree of affinity.

The remainder of this paper is structured or organized as follows. Section 1 provides the related work in this area. Section 2 describes our system called SNex and the extraction process. Section 3 is about the visualization of the social network. Section 4 presents the results and conclusions and, finally, Section 5 describes the future work.

### 1. Antecedents

The first system dedicated to the automatic extraction of information was the Referral Web by Kautz, Selman, and Shah [3] in 1997. This system was designed to extract social networks through the search engine Altavista. The social network generated was ego-centric, i.e. it was centred on a determined individual. The basic idea of Referral Web is based on the theory of the six degrees of separation and the list of potential individuals related to the original actor was refined using the Jaccard coefficient [4] to obtain the affinity calculations. The main problem was the ambiguity of the names and an attempt was made to resolve this by adding extra fields for the user to clarify names.

A more recent system (presented in the sector) is that of Flink de Mika [5], which provides an in-depth study of the conclusions of Referral Web and improves with additional functions. The architecture of Flink is divided into three layers: data acquisition, storage and visualisation. The system was capable of visualising social networks of semantic network communities. The data obtained was from the web, from the FOAF profiles, from emails and from published articles. The web-mining process employed by Flink is similar to that of the Referral Web and it also uses the Jaccard coefficient [4].

Further work worth mentioning is Polyphonet by Hamasaki and Matsuo [6], also based on [3]. The system works from a list of names by extracting the social network related to the data which the search engine Google provides. Unlike the previous systems mentioned, Polyphonet uses the Overlap coefficient to calculate the affinity between actors. Polyphonet also incorporates methods to detect the type of relation between individuals, which might include co-authoring, participation in joint projects, etc. It also tries to improve scalability by reducing the number of hits on the search engine in order to reduce the computational costs of the extracting system.

The work carried out by De la Rosa and Martinez Gasca [2], which analyses the results of Polyphonet and Flink, should also be highlighted. A methodology to extract social networks is presented where electronic mails of the hits on the search engine are used to construct the extraction process.

SNex, the system presented, is a continuation of their most recent lines of research. On the one hand, the extracting system is fed by the information obtained from the hits
on Google and Yahoo, based on the Polyphonet models. On the other hand, the base for information searches is based on the emails of the actors, as proposed by De la Rosa [2]. We must also consider the scalability parameters to optimise temporary costs. In the system, any of the coefficients described in the previously mentioned works can be employed to calculate affinities. Furthermore, the system incorporates new metrics which can be used to assess the affinity between actors by referring to the semantic descriptors extracted from the information obtained from Internet. This innovating aspect differentiates the system from the previously mentioned works.

2. Description of the System

The extraction process of a social network has three clearly differentiated phases:

1. Web crawling: tracks all the information available an actor. The crawler takes advantage of the capacity of existing search engines to carry out this process.
2. Web mining: extraction process, which uses the information gathered from Internet and the semantic content of each actor as its starting point. New actors linked to the original actor who will in turn expand the social network are obtained from this information
3. Visualisation: navigation tools which enable us to browse on the social web and to consult all the stored information.

The typology of the social networks (researchers who usually belong to a university) to be extracted has been restricted in this application, as have the search techniques and the type of content to be used in the web mining process.

2.1. Web Crawling

The first step when making a query on the search engines is to determine the access technique in order to carry out a consultation. Two search engines were used in this work: Google and Yahoo. The access to the search engine Yahoo was made through its API [7]. In the case of Google queries were made directly through the search engine and a parser was created with the results.

Once access to the search engines had been prepared, the next step was to make the queries. The ambiguity of the results of the searcher when it operates with names and surnames was affirmed. The previously mentioned studies, [5, 6], attempt to resolve this problem with different techniques. In the case of Polyphphonet, the name of the actor and keywords to differentiate one actor from another are added. However, given the research environment, this technique would not be considered a valid solution since, on one hand, it involves finding out information about the actor (something which is not always possible when we are dealing with an expansion of the social network) and, on the other, in tests carried out, the keywords used for the main actor did not always succeed in eliminating ambiguity.

One approach to resolve the problem is that described by De la Rosa [2] which uses the actor of the email as an identifier. The results obtained by the searcher in the tests carried out match the requested actor 100%.

Finally, an attempt was made to restrict the information from Internet to that which was strictly relevant to the social network and we aimed to reduce the results which do
not provide such information. Information of interest to researchers in the academia is found in articles and published work (academic papers). Information on the areas in which they develop their work and further information about the subjects themselves can be found in such articles.

This type of document can often be found as published works in Internet and is usually in PDF format due to its portability. At the moment this affirmation may well be true. However, it is nevertheless limited to the majority of subjects concerning Computer Science and not necessarily the case for the rest of the university community. It is expected that advances in Information and Communication Technologies (ICT) will make articles such as the aforementioned on any subject matter available in the near future. Therefore, taking advantage of the parameters of the search engines, the only information from PDF documents will be retrieved.

2.2. Web Mining

Data extraction of the actors is the most crucial point of the process. Posterior analysis to determine the social network and the ties between the different actors will depend on this stage. The objective is to obtain information on an actor (relation of the actor and his semantic description through keywords) from the results obtained by the searcher, from the list of URLs associated to the actor.

![Snex 1.0 configuration environment](image)

**Figure 1. Snex 1.0 configuration environment**

The **mining** process has different stages: social network expansion, extraction of the semantic descriptions of the actors and, finally, the affinity calculations between the different actors who form part of the social network. The decisions taken at each of these stages are described below.

2.2.1. Social Network Expansion

The primary objective is to be able to expand the social network of the actor who is being analysed. It is important to identify all the actors related to the individual to be tested at this point. To achieve this list of PDF documents related to the individual being tested which the searcher has sent back is referred to.
The first step is to transform the PDF file to a plain text document. The extraction of candidates related to the actor being tested then proceeds. To undertake this task, the decisions taken to identify the valid candidates on completion of the tests will be described:

- Every email which appears in the content of the document is converted into a potential tie with the actor being tested.
- Every email which appears in the same line as the actor being tested will be considered as being of greater strength, especially if they are co-authors (affinity value equals 1).
- The emails found near the email of the actor being questioned will be considered as a direct relation (affinity=1). An e-mail at a distance of \( n \) characters (configurable in SNex) from the actor being tested is considered as near.
- All the emails that belong to a document with a density of electronic mail superior to \( d \) (configurable in SNex) will be penalised and will be of a lesser importance when calculating the affinity with the actor being tested. In this case, the affinity is multiplied by 0.2.
- Finally, all the e-mails which appear in document containing more than \( m \) emails (configurable in SNex) will be rejected. This measure ensures that documents which contain sending lists which do not provide any relevant information are deleted.

The application of these rules provides us with a list of candidates to expand the social network towards the next level. The acceptance of these candidates as members of the network will also depend on the affinity calculations made.

2.2.2. Extraction of Semantic Descriptions

The method used to extract the semantic content related to the actor being tested employs well-known text mining procedures, based on the NLP-IR system (Natural Language Process - Information Recovery).

The first step is to carry out a filter of the document in which all words, symbols and characters not providing useful information are deleted.

The second stage is to delete all the stopwords (words which do not provide us with any information, such as prepositions, articles, etc.) from the filtered text. To achieve this, the document language must be known. A procedure based on the deletion of stopwords has been developed. Previously created lists of stopwords generated from the grammatical and linguistic study of [8] are used to count the number of stopwords in the document. The highest number of stopwords found determines the language of the document.

Finally, in the third phase, a stemmer (or grouper) is in charge of grouping together words based on their root. Once this process has been completed, the lemmatisation, i.e. the representation through a single term of all the inflected forms of a word [9] is carried out. One of the first stemming algorithms was developed in 1980 by Martin Porter [10], hence the name, “the Porter stemming algorithm.” In our system, we used two stemming algorithms for Spanish and the Porter Stemmer for English [11] and the stemming algorithm for Galician created by Penbad [12].
2.2.3. Affinity Calculations

Once a selection of candidates related to the actor being tested has been confirmed and
the semantic descriptions of the content of the associated documents extracted, it is
now time to decide whether the ties between the actors do really exist and if these ties
are strong enough to take into consideration.

The aforementioned works [5, 6], which analyse the number of documents which
link the different actors, have been used as a base to calculate the ratio capable of
quantifying the degree of affinity between actors. The possibility of calculating affinity
with the Jaccard coefficient [4], employed by Flink and Referral Web, the Overlap
coefficient, employed by Polphonet, as well as the Cosinus Coefficient have all been
contemplated in the system. The idea is to configure a weighted combination of the
different coefficients in order to obtain a more reliable affinity calculation, although at
present, the weights of the coefficients is still under testing.

The Jaccard coefficient gives the relation between the documents in which both
actors appear in reference to the number of documents written by one actor or the other
(whether of the actors). The Overlap coefficient relates the number of documents jointly
written by X and Y, between the minimum of the results produced by actors X and Y.

The Overlap coefficient compensates for the limits of the Jaccard coefficient which
occurs when author X has many documents while Y has few but all of Y’s documents
are linked to X. The Jaccard coefficient provides a value which is of little relevance.
The Overlap coefficient rectifies this deficiency.

The Cosinus coefficient resolves the deficiencies of the Overlap coefficient which
is produced when, for example, actor X has 100 documents and actor Y has only one
document but this document is shared with actor/author X. If we apply the Overlap
coefficient the affinity is maximum whereas with the Cosinus coefficient it is 0.1.

\[
\text{Jaccard} = \frac{n_{X \cap Y}}{n_{X \cup Y}} \quad \text{Overlap} = \frac{n_{X \cap Y}}{\min(n_X, n_Y)} \quad \text{Cos} = \frac{n_{X \cap Y}}{\sqrt{n_X \cdot n_Y}}
\]

From these affinity calculations and taking into account the criteria described in
section 2.2.1, the candidates to form part of the social network are filtered. To complete
this filter, an affinity value of 0.2 has been determined as a threshold.

Another type of affinity between actors is that which can be obtained through the
semantic descriptions extracted from each of the individuals who make up the social
network. This method enables us to link actors who may not share a high degree of
affinity but who do work in similar areas and consequently put them in touch with each
other.

To carry out this calculation, a list of keywords has been selected for each actor.
The keyword list is made up of keywords taken from their documents, selecting the
words with the highest occurrence in the text. In the SNex Version 1.0, the 10 most
frequent words are selected for each actor. This information enables us to develop the
search of a formula or metric which can be used to calculate the semantic affinity. The
keyword groups of each actor are defined to carry out this task.

\[
L_X = \{\text{stem}_0, \text{stem}_1, \ldots, \text{stem}_n\} \quad L_Y = \{\text{stem}_0, \text{stem}_1, \ldots, \text{stem}_n\}
\]
An initial formula has been developed to calculate semantic affinity. The theory behind it is based on the calculation of the number of keywords which coincide with both actors. The keywords are weighted according to their occurrence. The importance of a keyword depends on the number of times it appears in the content of an actor’s documents.

\[
\text{Affinity} = \frac{\sum_{i=1}^{n} \exists(L_{X_i}, L_{Y})}{n}
\]

The expression \( \exists(L_{X_i}, L_{Y}) \) gives 1 if the keyword \( i \) from the keyword list of actor \( X \) is found in the keyword list of actor \( Y \) or 0 if does not exist. Therefore, if two actors have 5 of the 10 keywords in common, their affinity will be 0.5.

Having analysed the tests carried out, it was concluded that a study of the importance of the word within the whole social network was required. If there is a high incidence of the word in the whole social network, this does not determine a greater degree of affinity between two actors as it is present in the whole network. Alternatively, if the word is common to two actors but not frequently used in the social network, this indicates that the connecting link is due to the correlation of the semantic interests of both actors which the rest of the social network do not share. Therefore the initial calculation which weighted this factor has been modified and the following formula obtained:

\[
\text{Affinity} = \frac{\sum_{i=1}^{n} \exists(L_{X_i}, L_{Y}) \cdot \frac{\text{Occurrences}(L_{X_i})}{\text{GlobalOccurrences}(L_{X_i})}}{n}
\]

By this means, if actors \( X \) and \( Y \) coincide in 5 out of the 10 keywords, but these words are commonly used in the social network, the number of global occurrences will be very high, which will make the factor reduce the affinity between actors.

### 3. Visualization

Although it is not the main focal point of this research, it is essential that the environment for the presentation and navigation of the extracted social network is friendly and usable. An environment in which results can be easily obtained will facilitate future analysis and corresponding conclusions. Three functionalities have been designed for this purpose:

- Visualiser of the graph of an actor: this provides us with a vision of the social network, showing the actor and the different ties he shares with other actors. The keywords associated with the actor are also shown. The implementation of this function is based on the work of Moritz Stefaner [13]
- Tag cloud: the most relevant keywords of the whole social network are shown. This allows for the filtering by the given words with matching actors. These words are then used in subsequent filters carried out on the actors
• Navigation panel: this enables us to analyse the various levels of the social network and to consult the affinity coefficients of each actor as well as providing information on the performance of the system in its searches and data extraction.

Figure 2. Visualisation of the social network

4. Results and Conclusions

Firstly, a comparative study of the response time of the two search engines on which SNex can currently consult information, Google and Yahoo, was carried out. The response time for 100 queries on the Yahoo engine was 85.620 seconds (0.86 s/consultation), while for Google it was 43.480 seconds. Here we can see how the response time of Google is much faster than that of Yahoo, almost half.

With reference to the method used to detect the language, tests on 81 PDF documents were carried out and the percentage of matches was 96%. In fact, on analysis of the three documents in which the language was wrongly identified, we saw how this error was due to a coding problem related to the PDF format. Therefore we can conclude that the detection method works correctly.

Turning to the reliability of the networks obtained by the automatic extraction process, the fact that the test environment consisted of queries made by research professors from our university centre should be underlined. The election of this environment permits us to verify the results obtained by consulting with those involved.

The first tests were carried out on known authors and the results obtained were absolutely correct. That is to say, the actors with a degree of co-authoring in the PDF documents were found. This meant that none of the actors were eliminated from the process given the fact that co-authoring implies a strong tie between actors.

Another of the tests carried out was realized with ralsina. This person was chosen due to her productivity in research articles. A level 1 network was initially created with 93 related actors. However, many of these actors had no real ties with the candidate. The reason for this as that co-authoring had not been penalised at this point and that sending lists from PDF documents had been classified as co-authors of an article. The results obtained were significant. From the 93 candidates, only 7 were actually actors with real ties. Applying the Overlap coefficient, 14 ties emerged, while when the
Cosinus coefficient was applied only 6 were found. The results obtained with cosinus are closer to the truth, if we consider that only one real tie was lost, while the Overlap coefficient produced 8 additional “unreal” actors related to the main actor.

Another noteworthy consideration is that the very nature of the semantic affinity formula is not commutative. For example, $\text{Affinity}(X,Y)$ does not necessarily have to be equal to $\text{Affinity}(Y,X)$. This has been proven in the tests. Although it is true that it might not seem logical at first, a series of tests demonstrated how the importance or strength of the tie between two actors does not give the same result when analysed from each of their different perspectives or roles within the social network. It may be the case that actor X belongs to a social network simply because he is linked to actor Y who forms part of the same network. In this case, the degree of affinity from actor X towards Y should be high. Nevertheless, the degree of affinity from actor Y towards X is not necessarily as relevant as in the previous relationship given that Y could be linked to other actors who are more important to him than actor X.

This application has been developed with the adoption of a vision-controller model which separates the layers of data, business logic and user interface. The SNex application has been developed using the PHP programming language and the MySQL database. With reference to system performance, the extraction system was extremely slow at first. To resolve this issue, **multithreading** was incorporated into the system so that several queries could be made on the search engines simultaneously. Response time on the search engines was reduced. The results were then stored in the memory cache and on hard disk to avoid redundant or repeated queries. This method resulted in a drastic time reduction. At present, it usually takes between 100 and 350 seconds to generate a level 1 social network.

5. Future Work

Focusing on SNex 1.0 and the results obtained, there are multiple future lines of research whose goal is to improving performance levels of the social network extraction system and to develop more precise methods to calculate affinity levels.

As far as web crawling is concerned, the possibility of testing the system with several additional search engines is gaining ground. The possibility of using the results obtained by the different search engines and assigning a weighted factor to each one is also being considered. Another aspect to take into account is how to improve the search techniques of the search engines.

The improvement strategies to work on the web mining stage are multiple. On one hand, proposals have been made to widen the scope of the type of information which is analysed. At present, only documents in PDF format are analysed but there are plans to study whether the addition of other types of document (html, postcript, text, etc.) would improve the quality of information and lead to better results in the social network.

In the area of extraction of semantic information, more complex algorithms are set to be included in this area in order to obtain semantic abstractions, such as the use of Wordnet at Princeton University.

Research and development into affinity calculations will continue and both coefficients and semantic descriptors will be employed. In this area, the investigative qualities of the researcher tend to be assessed and quantified by the number of references made by other authors in their articles.
Acknowledgements

The authors would like to thank Enginyeria i Arquitectura La Salle (Universitat Ramon Llull) for their encouragement and assistance, especially Lisa Kinnear for the linguistic reviews of the paper.

References