OSLa Simões, Barreiro, Santos, Sousa-Silva & Tagnin (eds.) *Linguística, Informática e Tradução: Mundos que se Cruzam*, Oslo Studies in Language 7(1), 2015. 183–207. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

AFFINITY MINING OF DOCUMENTS SETS VIA NETWORK ANALYSIS, KEYWORDS AND SUMMARIES

PAVEL BRAZDIL, LUÍS TRIGO, JOÃO CORDEIRO, RUI SARMENTO AND MOHAMMADREZA VALIZADEH

RESUMO

Encontrar pessoas com interesses semelhantes dentro de um domínio pode fornecer um importante auxílio na gestão de centros de investigação. Como a produção académica é facilmente obtida em bases de dados bibliográficas e académicas, estas podem ser usadas para descobrir as afinidades entre os investigadores que não estejam já evidenciadas pela co-autoria. Este processo de descoberta dá-se com a ajuda de técnicas de análise de texto, na base dos termos utilizados nos respectivos documentos. A afinidade pode ser representada em forma de rede, em que os nós representam os artigos de cada investigador e as ligações representam similaridade entre os diferentes investigadores. Cada nó pode ser caracterizado através de diversas medidas de centralidade na rede e algoritmos de detecção de comunidades permitem identificar grupos com interesses semelhantes. Cada nó é ainda caracterizado por um conjunto de palavras-chave e resumos descobertos automaticamente com a ajuda de técnicas avançadas. Este artigo fornece mais detalhes sobre os métodos adoptados e/ou desenvolvidos, alguns dos quais foram implementados no nosso protótipo. Os métodos descritos são gerais e aplicáveis a muitos domínios diferentes, incluindo documentos que descrevem projetos de I&D, documentos associados a legislação, processos judiciais ou procedimentos médicos. Acreditamos deste modo que este trabalho pode ser útil para um público relativamente amplo.

[1] INTRODUCTION

Researchers seek to discover other researchers with similar interests to follow their work and plan future collaborations. At management level, this knowledge enables to identify suitable researchers for a given task, which precedes the implementation of partnerships with other institutions and researchers policies. Another advantage of this analysis is that it goes beyond the formal hierarchical framework within the organization, thereby revealing its unknown connections that can be followed up.

The main scientific contribution is beyond re-using standard techniques of text mining to bibliographic databases, but rather using these techniques to ob-

tain two kinds of graphs, co-authorship and affinity graphs, and exploring a differential analysis with the aim of identifying new useful knowledge.

Our aim is to focus on affinity analysis between certain R&D centers for various reasons: First, the outcome of the study may be useful to these centers. It may propose that certain collaborations be initiated. Besides, the outcomes of automatic analysis may be easily verified by some members of these centers. This research could be extended later to cover a larger set of centers.

Regarding the particular bibliographic database, we have chosen *Authenticus*[1], as its design is based on Bugla's work and was adopted by the University of Porto. It has the advantage that it retrieves publications from several other bibliographic databases (incl. e.g. SCOPUS).

Regarding the discovery of similarities between researchers, Price et al. (2010) developed a methodology for the Web, called *SubSift*, which enabled to establish profiles for researchers on the basis of researchers' publications. Based on these profiles, a typical Information Retrieval task is performed aiming to compare the papers submitted to a scientific conference (playing the role of Query in IR) with different profiles, in order to optimize the task of distributing articles to review.

A similar idea was followed by Trigo & Brazdil (2014), although the aim in this work was different — to uncover affinity among researchers that are not evidenced by co-authorship. Various researchers have analyzed co-authorship networks (e.g. Bugla 2009; Choobdar et al. 2012), but no one to our best knowledge has analyzed the differences between the two types of information. To uncover these we resort to many diverse techniques.

The publications titles are extracted into plain text files, each representing a particular author. The text files are retrieved and preprocessed in the usual manner. We use *bag-of-words* (BoW) and *vector* representation (Feldman & Sanger 2007), but perform usual preprocessing including removal of numbers, stop-words, punctuation and other spurious elements. After this task, the list of documents is transformed into a document-term vector representation with *tf-idf* weighting. The vector representation is used to generate the cosine similarity matrix. This matrix can be visualized in the form of a graph and is used as the basis for further processing following (Iacobucci 1994).

After transforming the similarity matrix into a graph format, we use the community discovery algorithm. There are many approaches that could be used for this aim. Here we mention one of them — *Walktrap* (Pons & Latapy 2006). This technique finds densely connected sub-graphs, also referred to as communities, through random walks. It assumes that short random walks tend to stay in the same community.

[1] Authenticus bibliographic database. https://authenticus.up.pt/.

The affinity network enables to calculate certain measures of importance of the researchers within their affinity group and in the context of different communities. This involves different centrality measures (Wasserman & Faust 1994). The *degree centrality* is based on the number of connections to a vertex. The *betweenness centrality* indicates the number of times a vertex joins two other vertices on the shortest path. The *eigenvector centrality* shows the importance of vertices that connect to a given vertex. Some centrality measures can be computed to account for different weights of the connections.

Our work is also concerned with the problem of characterizing each individual/subgroup with appropriate keywords or short summaries. This is important, as the user does not only want to identify individual subgroups, but also see what distinguishes them. In terms of distinction, we can consider techniques from forensic linguistic analysis (Sousa-Silva et al. 2010) to better shape the subgroup/author textual boundaries. As for automatic keyword generation, there are really many approaches that could be followed. In our previous studies we have explored the approach of TextRank (Mihalcea & Tarau 2004), who used a graph-based language independent key phrase extractor. They explored the fact that many multi-word units can be identified by looking at relative positions in which these occur. This is because there is a tendency for a pair of single-words to co-occur in fixed positions relatively to each other. We plan to evaluate different approaches in the future and adopt the one that achieves the best results.

Besides keywords, users / documents can also be characterized using automatically generated summaries. Automatic text summarization (ATS) aims at the transformation of textual information into a more humanly tractable representation. Normally, this transformation involves a reduction of the original text by eliminating the irrelevant portions, while maintaining the most relevant ones.

In this area, a great number of methods have been experimented throughout the last twenty years, following mainly extractive approaches (EA) (Erkan & Radev 2004; Wei et al. 2008; Valizadeh & Brazdil 2015, 2014), which basically summarizes texts by selecting the most relevant sentences. One rather successful approach uses supervised learning to do this (Valizadeh & Brazdil 2015). It tries to enhance the coherence of the summary by trying to detect a particular kind of anaphoric chain — actor-object relationship (AOR) between sentences. The sentences that satisfy this relationship have their importance value enhanced.

Extractive approaches have the disadvantage that they permit the inclusion of rather long sentences into the summary. We have tried to overcome this by generating shortened versions of such sentences with the help of machine learning methods (Cordeiro et al. 2013). However, other possible transformations / reformulations could be considered in the future. We intend to explore them in collaboration with linguists from the Faculty of Arts of the University of Porto (FLUP).

[2] METHODOLOGY

This section presents the main steps undertaken to uncover the unknown information regarding affinities. The method involves the following steps:

(i) Identify institutions and obtain researchers' names;

(ii) Use web/text mining to process researchers' publications;

(iii) Elaboration of similarity matrix and visualization as a graph;

(iv) Discovering potential communities linked by affinities;

(v) Elaboration of a co-authorship graph and differential analysis of graphs;

(vi) Identification of important nodes (researchers) in the graph;

(vii) Characterization of nodes using keywords;

The details about all these steps are given in the following sub-sections. Additional functionalities that are not part of the implemented prototype include:

(i) Characterization of nodes using summaries;

(ii) Learning to generate shortened sentences for summaries.

The details about all these steps are given in the section [3].

[2.1] *Identify institutions and obtain researchers' names*

Our pilot study was carried out in conjunction with a dataset that includes approximately 3000 publications of about 100 researchers belonging to 5 different R&D centers of INESC Tec (LIAAD, CRACS, CESE, CTM and CEGI). This data was provided by the authors of Authenticus database discussed further on. Therefore we did not require any sophisticated procedure to obtain this. However, in general, it may be necessary to retrieve this information from websites and so in the rest of this section we describe the method.

Each research institution has normally a webpage listing their researchers. Lists of researchers can be extracted by building an expression in the *XPath* query language to obtain their names from the website. Regarding implementation, different languages can be used. We have used R and exploited its *tm* package for part of text mining and the *XML* package for web mining.

Each researcher's name can be used in the search through the chosen bibliographic database, such as DBLP, which enables direct access to each researcher list of publications. The retrieval of publications can be done automatically, using *XPath* expressions. However, a problem of *named entity identification* arises here. Typically, one of the variants will appear on the institution site, which may

not match the name used in the bibliographic database. Also, as researchers may have several variants of their name, several entries may exist in the bibliographic database for the same researcher. So these issues need to be resolved.

It could be argued that the researchers' names might not be retrieved from the web pages of a particular research institution / R&D center, as these appear in the articles. This approach has, however, a disadvantage that the set of research institution / R&D centers would grow, as more articles would be encountered and processed. We prefer to restrict the number of R&D centers to a certain pre-defined set.

Another problem is that we may have several investigators with the same name in the bibliographic database. One of the techniques used by Bugla (2009) is the following. To determine whether a given publication of P in some bibliographic database should be attributed to person P' on a given site, a check is made whether both (i.e. P and P') have the same home institution. Various other researchers have investigated the issue of determining whether several variants of one name belong to the same author and various methods have been proposed (e.g. Santos & Ribeiro 2011).

Regarding the particular bibliographic database, we have chosen Authenticus database, which was developed by the University of Porto, because it retrieves publications from several other bibliographic databases (incl. SCOPUS, Google Scholar, ISI Web of Science, DBLP and Orcid). In the work reported here, we were able to skip many of the Web/Text Mining steps just described, as we were provided with a database that included all relevant information.

[2.2] *Use web/text mining to process researchers' publications*

The publications titles are extracted into plain text files, each representing a particular author. The text files are retrieved and preprocessed in the usual manner. We have used BoW representation, removed numbers, stop-words, punctuation and other spurious elements. After this task, the list of documents is transformed into a document-term vector representation with tf-idf weighting (Feldman & Sanger 2007).

[2.3] *Elaboration of similarity matrix and visualization as a graph*

The vector representation described in the previous step is used to generate the cosine similarity matrix. This matrix can be visualized in the form of a graph and is used as the basis for further processing following (Iacobucci 1994). Figure 1 shows an example of an affinity graph for the R&D unit LIAAD. Each researcher is represented by a circle and its size is related to his/her number of publications in the Authenticus database. The thickness of the edges represents the similarity value between pairs of researchers. The wider the line, the more similar the two researchers are joined by this link. For simplicity all links / similarities below a

given threshold have been considered irrelevant and removed. The value of the threshold was chosen somewhat arbitrarily, but in future the user will be given an option to adjust it according to his/her needs. Besides, we note that the software determined the length of each edge automatically. The value of similarity is taken into account in this process. In general, the nodes with high similarity appear closer than others.

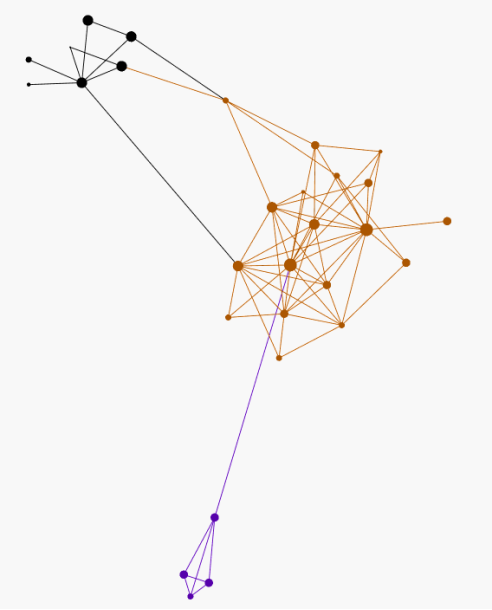

FIGURE 1: Researcher affinity network for R&D center LIAAD of INESC Tec

Visualization tools

Visualization tools play an important role in data analysis, as visual information organization enables the analyst/user to interpret and detect patterns or other relevant information faster and more effectively. This requires developing tools that show the information in an intuitive and interactive way.

The developed web application — prototype Affinity Miner[2] — is based primarily on R language and an appropriate set of packages. With the data conveniently indexed we used R as a language platform for the implementation and to represent the data which needs to be conveniently indexed.

For this task we use the shiny package (RStudio, Inc 2014) that is a web application framework. In this way, our web application can react instantly to user inputs with the goal of changing the output displayed to the user.

[2] See http://gallicyadas.pt/affinity-miner/.

Another requirement is the output availability in remote locations and the use of standardized frameworks and software (e.g. HTML, JavaScript etc.). The best way of doing this is by presenting the output, including network graphs, in a web browser. For this task we chose *sigma.js* library, a JavaScript library dedicated to graph drawing (Jacomy 2013). It enables the network display on web pages and may be used to integrate network exploration in rich web applications.

[2.4] *Discovering potential communities linked by affinities*

After transforming the similarity matrix into a graph format, we use the community discovery algorithm called Walktrap (Pons & Latapy 2006). This technique finds densely connected sub-graphs, also defined as communities, through random walks. It assumes that short random walks tend to stay in the same community.

The hierarchical agglomerative approach is based on a measure of distance between vertices (node to node). An optimal level of modularity of the network, based on the weighted connections between internal and external community is used by the algorithm to identify non-hierarchic communities.

Different discovered communities can be identified in the graph in different ways. In our case we have used color to highlight different communities. In the example shown in Figure 1 the system discovered three communities. The largest one on the right (ochre), corresponds to machine learning / data mining subgroup. The one at the bottom of the figure (purple) corresponds the subgroup working in optimization and operational research. Finally, the one at the top (black) includes researchers working in mathematical modeling.

The graph elaborated on the basis of the five R&D centers of INESC Tec (Figure 2) is much more complex than the one shown on Figure 1. It includes 13 communities, some larger, others smaller. The machine learning / data mining affinity group is identified by blue color. As can be seen, it includes quite a relatively high number of researchers belonging not only to LIAAD, but also to CRACS, CESE, CTM and CESI. This information is available in our prototype.[3]

The communities discovered may not correspond to the organizational structure of the given institution. Such situation can be analyzed to determine whether this is desirable, or else what would be the best organization to consider.

[2.5] *Elaboration of a co-authorship graph and differential analysis of graphs*

The generation of the co-authorship graph is a relatively simple matter. A link between authors A_i and A_j is introduced, if they are co-authors of at least one of the papers. After the affinity and co-authorship graphs have been constructed, it is possible to proceed to carry out a *differential analysis*, following Trigo & Brazdil

[3] Available from `http://gallicyadas.pt/affinity-miner/`.

FIGURE 2: Researcher affinity network for the 5 R&D centers of INESC Tec and identified communities

(2014). This involves constructing a graph that represents basically the difference between the two graphs.

The following two figures illustrate this. Figure 3 shows a part of co-authorship graph that includes some researchers of LIAAD. Figure 4 shows a part of *differential graph* resulting from the differential analysis. It shows all the affinity links that do not have a corresponding link in the co-authorship graph.

For example, we note that Márcia Oliveira has just one co-authorship link — with João Gama, while the differential graph shows three other affinity links — to Alípio Jorge, Pedro Campos and Pedro Quelhas Brito. These links have been revealed by the differential analysis. Such links may be of interest firstly to the researchers involved, but also to the management when creating new teams for a new project.

[2.6] *Identification of important nodes (researchers) in the graph*

The affinity network enables to calculate certain measures of importance of the researchers within their affinity group and in the context of different communities. This involves different centrality measures (Wasserman & Faust 1994). *Degree centrality* is based on the number of connections to a vertex.

Betweenness centrality indicates the number of times a vertex joins two other vertices on the shortest path. A node with high betweenness centrality has a large influence on the transfer of items through the network, provided that the transfer follows the shortest paths.

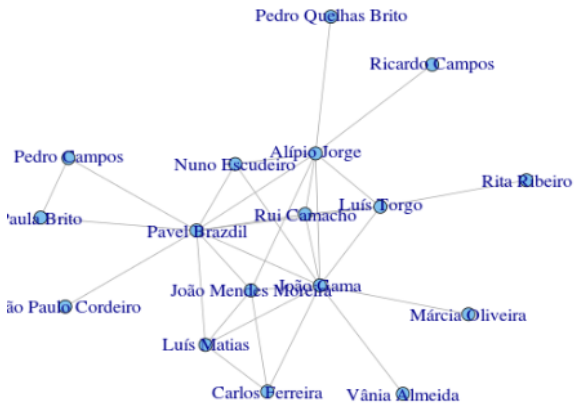

FIGURE 3: A part of co-authorship graph for LIAAD

FIGURE 4: A part of differential graph (affinity — co-authorship) for LIAAD

Eigenvector centrality shows the importance of vertices that connect to a given vertex. Some centrality measures take into account different weights of the connections.

Closeness can be regarded as a measure of how long it will take to spread information from a given node to all other nodes sequentially. It can be calculated on the basis of *farness* of a given node. This measure is defined as the sum of its distances to all other nodes. Closeness is defined as the reciprocal of the farness.

Currently, our prototype shows two of the centrality measures for a chosen researcher (betweenness and eigenvector centrality). The centrality values are not significant in themselves, but need to be compared to other ones. Thus we can affirm that a particular researcher has rather high betweenness centrality, when this value is high in relation to others.

[2.7] *Characterization of nodes using keywords*

In general, it is important to have a concise description for each node that will provide a quick overview of the content of that node. The issue of characterizing each node with appropriate keywords enables the user to decide whether to zoom on this node while searching for relevant information, or otherwise focus on some other part of the network.

As for automatic keyword generation, there are really many approaches that could be followed. Up to now we have explored the approach of Mihalcea & Tarau (2004), who used a graph-based key phrase extractor. This approach is incorporated in our prototype. The *TextRank* is a multi-word unit extraction algorithm that explores the *centrality measure* of *PageRank*. The latter algorithm infers the importance of a web page by the number of web pages that have links to it, while taking also into account their relevance measure. Rather than connected pages, TextRank considers adjacent terms that may also be represented as a graph. Terms are represented by nodes and undirected edges represent their co-occurrence. Following the approach of *PageRank*, the multi-word terms that are similar to others are used to increment their importance.

The *DegExt* algorithm (Litvak et al. 2011) is also based on a graph-based representation, but takes into account the order of terms in the process of constructing directed graphs. It is assumed that the more often the terms appear linked, the more relevant they are.

Ventura (Ventura 2014; Ventura & Silva 2013) explored the fact that many multi-word units can be identified by looking at relative positions in which these occur. This is because there is a tendency for a pair of single words to co-occur in fixed positions relatively to each other. So, for instance, the multi-word unit *surgical abortion* can be identified, as the term *surgical* co-occurs quite often in 1 position to the left from the term *abortion*.

We plan to re-evaluate different approaches and adopt the one that achieves the best results.

Evaluation of keywords generated

The present quality of the characterizing keywords generated by our prototype is quite reasonable, at least when considering some of the terms generated. So far, we have performed an informal evaluation, by just comparing the keywords generated with the keywords extracted from researchers' web pages. We plan to carry out a more thorough quantitative study later using conventional term overlap metrics (e.g. *precision*, *recall*).

Let us consider some examples shown in Table 1, in which we compare the real profile (**R**) expressions, characterizing two researchers from LIAAD, with the list of key words automatically generated (**S**) by our current prototype.

Researcher		Keywords
Pavel Brazdil	**R**	***Data Mining*** *and Decision Support; Algorithm Selection via Metalearning and Planning; Meta-Learning; Web Mining, Text Mining and Web Intelligence;* ***Artificial Intelligence****.*
	S	classification algorithm; logic programming; inductive logic programming; knowledge discovery; data mining; artificial intelligence;
João Gama	**R**	***Data Mining*** *and* ***Decision Support****; Knowledge Discovery from* ***Data Streams****;* ***Artificial intelligence***
	S	data stream; decision tree; change detection; knowledge discovery; data mining; sensor network; artificial intelligence; classification algorithm; computer science; sensor data; decision support system;

TABLE 1: Comparison of the automatically selected keywords (**S**) with their real keywords (**R**) obtained from web pages

Table 1 shows that several keywords agree well with the real ones, identified by the researchers on their web pages. It appears that the real expressions are more meaningful and would lead to better thematic assessment. In this area it is important to avoid both too general keywords (e.g. computer science) and too specific ones. This reveals the need for further studies in this area, which is related to the problem of summarization using short sentences or snippets discussed next.

[3] CHARACTERIZATION OF NODES (RESEARCHERS) USING SUMMARIES

[3.1] *Extractive summarization of nodes (researchers)*

Nodes (representing individuals with associated sets of documents) can also be characterized using automatically generated summaries. Automatic text summarization (ATS) aims at the transformation of textual information into a more humanly tractable representation. Normally, this transformation involves a reduction of the original text by eliminating the irrelevant portions, while maintaining the most relevant ones. This approach is referred to as *extractive summarization*. An alternative to this is *abstractive summarization* whose aim is to produce a short text and the process can include reformulation of the given set of sentences. In this section we will focus on extractive summarization. The methods can be divided into unsupervised approaches and supervised ones. Both will be discussed in the following in some detail.

Unsupervised graph-based approaches

One key work in this area is that of Erkan & Radev (2004) who proposed a graph-based method, referred to LexRank, where nodes represented sentences and links between two sentences the measure of similarity between them. PageRank algorithm was adopted to enhance the importance of nodes (sentences). Thus a node (sentence) that is similar to many other important nodes (sentences) is likely to end up with a high score. Sentences were then selected according to the score. The main steps of this approach are shown in Figure 5.

FIGURE 5: Basis steps in unsupervised graph-based summarization

Otterbacher et al. (2005) adapted LexRank algorithm to topic-sensitive multi-document summarization. This algorithm is known as T-LexRank. Wan et al. (2006) proposed a topic-sensitive graph-based model that was used for a query-based multi-document summarization. They used two graphs to show inter- and intra-links in query-oriented multi-document summarization.

Wei et al. (2008) extended the previous work in two aspects: First, by using the *centroid* value of words in the algorithm for generic summarization task and second, by exploiting similarity between documents in query-based multi-document summarization task. They showed that their algorithms, DsR-G and DsR-Q, lead to better summaries than earlier approaches.

We have improved these algorithms further (Valizadeh & Brazdil 2015). The improvement was mainly due to the inclusion of the concept of *density* (as an alternative to centroid) to the sentence ranking method. This was done both for generic and query-based multi-document summarization. The resulting algorithms, DensGSD and DensQSD lead to further improvements of summaries, as judged by the ROUGE measure (Lin 2004).

Supervised approaches for summarization

The aim of both unsupervised and supervised approaches is to generate a score for each sentence. The score is used to rank the given sentences and the sentences with the highest score are selected for the summary. However, the approaches differ in the way they calculate the score. The unsupervised graph-based approaches derive the score from a graph. The supervised methods used training data to construct a model with the help of machine learning (ML) methods. The model is used to predict the score (Toutanova et al. 2007; Ouyang et al. 2011; Valizadeh & Brazdil 2015).

The training data for supervised methods is in the form of a list of sentences $S_1 \ldots S_m$, each characterized by a set of n features and a score, which represents the target variable.

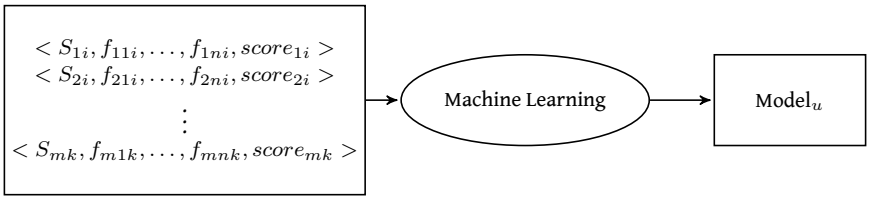

FIGURE 6: Training data for creating a model for a given document set DS

This scheme is illustrated in Figure 6. The index *i* (or *k*) represents a particular document, index *u* a particular human summarizer who has supplied the golden standard summaries.

Various features were proposed in the past. The features of Ouyang et al. (2011) were *sentence length without stop-words*, *sentence position*, *average tf-idf*, *sentence to query similarity*, among others.

Valizadeh & Brazdil (2014) enriched this set with some features that were derived from the graph-based representation, such as *sum of similarities between current sentence and other sentences*, *number of nonzero links sentence rank of T-LexRank*, besides others which lead to marked improvements in the quality of summaries.

Enhancing the coherence of summaries by detecting actor-object relationship (AOR) between sentences

Ideally, the sentences selected into the summary based on their scores should be coherent and supplement each other in their meaning. One method that can model this is by detecting a special case of direct anaphora, which was studied by Valizadeh & Brazdil (2015). This occurs when one sentence introduces an object that plays the role of an actor or a subject in another sentence. This relationship is referred to shortly as *actor-object relationship (AOR)*. The sentences that satisfy this relationship have their score value enhanced.

To be able to do this, it is necessary to use a parser. The authors have opted for the Stanford dependency parser, as it is freely available (de Marneffe et al. 2006). The parser returns, for each sentence, a set of relations of the type $tag(t_i, t_j)$, where *tag* characterizes the relationship between the terms t_i and t_j. The tags that were exploited in this work were, for instance, *dobj*, representing the direct object of the verb, $nsubj(t_j, t_k)$, representing a nominal subject/actor of the action. One example of a tag is $dobj(seize - 47, compound - 51)$. The items 47 and 51 represent identifiers determining where the words *seize* and *compound* appear in the parse tree.

The summary (SS) is generated sentence by sentence from the candidate ranked list (CS) for the test document set. The sentence with the highest score is selected for the summary (SS). After this, the combined ranked list (CS) is updated taking into account the sentence chosen for SS, AOR and MMR. Here, MMR represents the Maximum Marginal Relevance approach described by others (Carbonell & Goldstein 1998). Figure 7 illustrates how the AOR relationship is detected. It shows that the *object* of sentence number 3 in the summary is the *nominal subject* of sentences number 1 and 4 in the ranked list CS. Consequently, the scores of these sentences are increased.

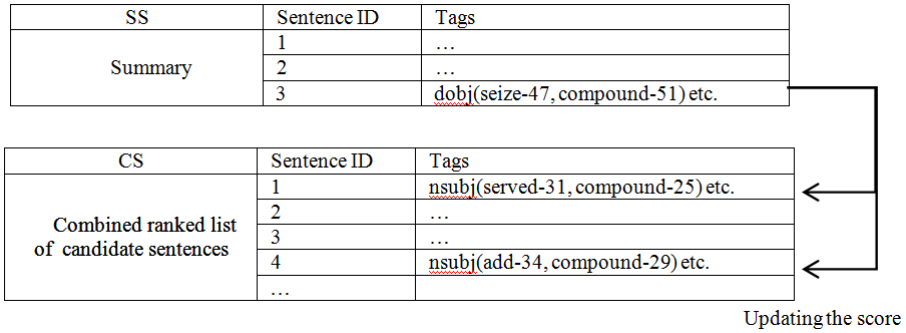

FIGURE 7: Detecting certain dependency parser patterns

If the AOR relationship has been detected, the corresponding score of the sentence in the candidate ranked list (CS) is increased by

$$\theta(Score(S_{highscore}) - Score(S_i))$$

where, $Score(S_i)$ denotes the score of sentence S_i in the candidate ranked list *CS*, $Score(S_{highscore})$ the score of the selected sentence for the summary (i.e. in SS) in the previous round. Moreover, θ is a parameter which determines the influence of this rule. Setting it to a high value means that the user would like AOR to have a strong effect on the sentence selection. The increased score of sentence S_i will increase the chance that this sentence will be selected into the summary.

After updating the sentence scores, the highest scored sentence in the candidate ranked list CS is selected to be included in the summary. This process is continued until the length limitation of the summary has been reached.

Valizadeh & Brazdil (2015) confirmed that this approach improves the quality of summaries significantly, as judged by the ROUGE values. As was shown here, this method enables to detect certain cases of direct anaphora, enhancing thus coherence between pairs of sentences in the summary. It is thus not surprising that this has a positive effect on the ROUGE score. ROUGE compares the generated

summary with human summaries and the latter tend to be more coherent than the ones generated previously by automatic methods.

[3.2] *Learning to generate shortened versions of sentences*

The art of being concise requires the ability to communicate ideas through a very short representation. In textual communication, this means not only to use fewer sentences but also choose the simpler and shortest ones. The aim is to achieve maximum efficiency. One particular line is concerned with characterizing texts with short snippets, that is, parts of the original sentences. Snippets are not so different to multi-word keywords. These can be obtained from the original sentences through various methods, like sentence decomposition and reduction (Cordeiro et al. 2013).

In the past decade, a number of works has been carried out in the field of sentence reduction. There is the work of Knight & Marcu (2002) who applied two machine-learning methods — a Bayesian model (noisy channel) and Decision Tree based model. This work was taken further by Galley & McKeown (2007) who explored probabilistic synchronous context free grammars (CFG). Clarke & Lapata (2006) proposed a hybrid system, where the sentence compression task is defined as an optimization of an integer-programming problem. Despite the fact that it is an unsupervised approach, it is completely knowledge driven, by a set of hand-crafted rules and heuristics that are incorporated to solve the optimization problem.

More recently Cohn & Lapata (2008, 2009) addressed a more complex issue of abstractive sentence compression/transformation by using a discriminative tree-to-tree transduction model, through a supervised learning setting. This work brings in new directions to the field, but still relies on supervised learning and deep linguistic analysis.

All of the above approaches rely on supervised learning or inclusion of manual knowledge. This is obviously a disadvantage. Normally, a training set of sentence reduction cases, manually selected and/or hand-crafted, is used, which is limiting in terms of scalability and applicability. Cordeiro et al. (2013) have pioneered a new approach were the training data is automatically collected from texts available on the web. Their aim was to develop an *unsupervised scalable methodology* for learning sentence reduction rules. In this work three important assumptions were made: (1) Only word deletions are possible and no substitutions or insertions allowed; (2) The word order is fixed; (3) The scope of sentence compression is limited to isolated sentences and the textual context is not taken into account. In other words, the compressed sentence must be a subsequence of words of the source sentence, which should retain the most important information and remain grammatical.

The methodology is based on a pipeline shown in Figure 6. First, some news sites are crawled with the aim of retrieving news stories about a certain topic. The news items are clustered by topic. The next step involves alignment and extraction of paraphrases, using text surface similarity measures (Cordeiro et al. 2007b) and specific alignment algorithms (Cordeiro et al. 2007a). Then pairs of paraphrases, are transformed into first order logic clauses, additionally enriched with certain linguistic knowledge. An example of a pair of paraphrases is shown in Figure 9.

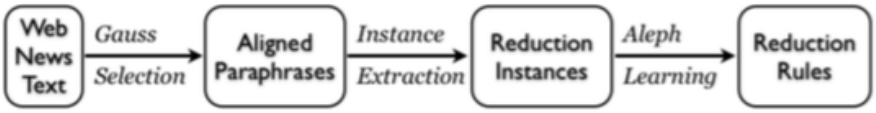

FIGURE 8: The pipeline architecture for learning sentence reduction rules from web news text.

```
Pressure will rise for congress to enact a massive fiscal stimulus package
Pressure will rise ___ ________ to enact a _______ fiscal ________ package
```

FIGURE 9: Example of a paraphrasic sentence pair, automatically extracted and aligned.

The massive corpus of aligned paraphrases is used to generate sentence reduction rules, with the help of a specific machine learning algorithm. The authors have opted for an *Inductive Logic Programming (ILP) system Aleph* (Srinivasan 2004). In this process, a combination of lexical and syntactical features is exploited: word tokens, part-of-speech tags, and phrase tags. For the syntactical tags, the Penn Treebank tag set was used. Figure 10 shows a pair of sentence reduction cases enriched by additional tags.

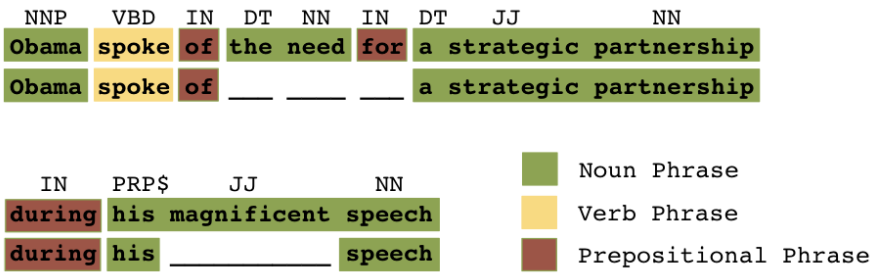

FIGURE 10: Two sentence reduction cases with three kinds of features highlighted.

The learning process yields a relatively large set of reduction rules which can then be applied to new sentences. A combination and even a composition of several reduction rules can be applied to a single sentence. The reduction rules incorporate different conditions, like for example, a restriction on the length of the eliminated segment. Besides, the reduced version should still maintain its grammaticality. For the former we use statistical lexical and syntactical models, automatically constructed from corpora. Example of two rules generated are shown in Figure 11.

Text Case	Rule
During his ~~magnificent~~ speech, the president ~~remarkably~~ praised IBM research.	L_1=PRP$ ∧ X_1=jj ∧ R_1=NN ∧ |X|=1 L_1=NN ∧ X_1=rb ∧ R_1=vbd ∧ |X|=1
He rallied support for an auto bailout and the ~~massive economic~~ stimulus he announced this weekend.	L_1=the ∧ X_c=NP ∧ R_1=NN ∧ |X|=2

FIGURE 11: Application of two learned reduction rules on two sentences.

The rules shown in Figure 11 include a conjunction of conditions, representing lexical and syntactical constraints for a given sub-sentence segment. The *X* letter represents the candidate elimination segment, while *L* and *R* represent the left and right position relative to X. For instance, in the second example, the rule expresses the following: *Eliminate a noun phrase segment (NP) of length 2 that is preceded by the word "the" and followed by a singular noun (NN).*

An effective automatic summarization system that incorporates sentence reduction can serve as a useful tool for creating small summaries characterizing the individuals and subgroups in the Affinity Miner. So far, different summarization systems exist as stand-alone prototypes. We plan to incorporate them in the Affinity Miner.

Summaries can be constructed from a collection of multi-documents, where a small and representative set of relevant sentences have been selected by the summarizer and subsequently simplified through our set reduction rules. This is quite important since the available space for characterizing nodes using summaries is limited, due to visualization space constraints. Therefore, the sentence reduction process allows us to incorporate a larger number of original sentences or snippets, yielding summaries with a higher information density.

As an example suppose that our system selects the following four sentences as the most relevant ones, describing a certain individual (researcher):

S_1: *In this work we investigate several new mathematical models for Plagiarism Detection.*

S_2: *As a conclusion, we have proposed a completely new algorithm on probabilistic topic modeling.*

S_3: *Our concern was to prove that LDA is the best-known approach for text segmentation.*

S_4: *This comparative study sets a new milestone in social network mining.*

Let us also assume that in order to satisfy size constraints, a limit of 25 words has been imposed on the length of the summary. As a consequence, we are only allowed to include two sentences, due to this limit. However, the use of a sentence reduction rule set could transform the original sentences in their reduced versions. We note that some of these (S_2, S_3 and S_4) are in the form of snippets.

S'_1: *We investigate models for Plagiarism Detection.*

S'_2: *A new algorithm on probabilistic topic modeling.*

S'_3: *LDA for text segmentation.*

S'_4: *A milestone in social network mining.*

This transformation allows us to display most of the information contained in S_1 - S_4, as S'_1 - $S_4{}'$ does not exceed the limit of 25 words. We note that each item in this list characterizes one particular area of research.

[4] CONCLUSIONS AND FUTURE WORK

Conclusions

We have presented a framework that uncovers research communities, real or potential ones, based on their scientific production. This is done by retrieving publication tiles for a given set of researchers, representing them in corresponding text files and elaborating a similarity matrix. This in turn can be used to construct a network of affinities.

Further processing leads to representations in the form of graphs. The community detection algorithms are used to uncover sub-graphs representing real or potential communities. These can be compared to the formal organization structure.

In our prototype we have devoted a special attention to the visualization of the graph of communities, as well as the characterization of its nodes (researchers). For this we have reused existing automatic techniques for selecting relevant keywords from texts.

Further steps involve differential analysis based on the affinity and co-authorship graphs. This analysis enables us to identify people that could potentially benefit from working together.

Future work

In the future we intend to process the abstracts or even full articles. We will consider also a substantially higher number of research centers and include thus more researchers. This represents some challenges for the process of elaborating the similarity matrix and the corresponding network. To overcome these, we plan to use the incremental / data-streaming approaches (Gama 2010).

It would also be interesting / useful to incorporate into our prototype certain techniques of *update summarization* explored recently by Costa (2014) who is a member of our group. This would enable to determine in what way a particular node differs from others.

As was shown earlier our current prototype is capable of characterizing each node with a set of keywords. In sections [3.1] and [3.2] we have discussed some aspects of our research in the area of automatic summarization. So far, these techniques have been implemented in the form of stand-alone programs. We plan to incorporate them in our prototype (Affinity Miner). This will lead to a more comprehensive characterization of nodes (researchers), which may be of interest to users.

A validation step needs to be added to our methodology. We plan to carry out a survey by questioning some researchers included in our study. We will inquire about who are the closest colleagues that conduct the most similar research. The outcome will be compared to the predictions obtained from the graph generated by our system.

An important issue that could be addressed stems from the fact that different researchers may use different vocabulary/terminology to describe the same concepts. This happens frequently when the researchers belong to different communities. This problem is difficult to overcome. It is possible to use, as some others did, *Wordnet* and *DBpedia* (Leal et al. 2012) to identify synonyms and related terms. This may be difficult for some specific domains, which may require the use of specific dictionaries, or the use of techniques that can identify potential synonyms (e.g. Grigonytė et al. 2010).

Another line of research that will be followed will exploit linguistic knowledge. We note that the sentence reduction can be attained through the transformation of an adverbial finite clause into a prepositional or adverbial phrase or non-finite clauses. Consider, for instance *"quando anoiteceu"* $\Longrightarrow$ *"à noite"*. In this example, the number of words is the same, the number of characters has been reduced, yielding a simpler and equivalent expression. Another example is the transformation of *relative clause* into a *gerundive* or *participial clause* (e.g. *"as garrafas que continham cerveja"* $\Longrightarrow$ *"as garrafas contendo cerveja"*). Since the same relations of meaning can be inferred in different types of structures, it is possible to use shorter sequences to convey the same meaning as the longer ones. To account for different semantic values of sentences, we will use a theoretical framework

that includes *rhetorical relations* (i.e. relations of meaning) (Asher & Lascarides 2003). The work on this line will build on the expertise of linguists from FLUP described in various publications (Silvano 2010; Leal 2011; Silvano 2012). This is a new and promising approach, as it joins researchers from two rather different research areas.

Regarding further management needs, we intend to go beyond similarity analysis with the aim to identify who should be collaborating with whom, considering their complementary capabilities / skills for a given task.

ACKNOWLEDGMENTS

This work has been partially funded by FCT/MEC through PIDDAC and ERDF/ON2 within project NORTE-07-0124-FEDER-000059 and 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 FCOMP-01-0124-FEDER-037281.

We wish to thank Fernando Silva and his collaborators, who are responsible for the Authenticus bibliographic database, for providing us with data that we needed for this study — titles of publications of INESC Tec researchers.

We wish to thank also the colleagues working from FLUP carrying out research in the area of linguistics — Fátima Oliveira, M. da Purificação Silvano and António Leal — for taking interest in abstractive summarization and their willingness to contribute. This may open possibilities for interesting new advances in the future.

REFERENCES

Asher, Nicholas & Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Bugla, Sylwia. 2009. *Name identification in scientific publications*. University of Porto MSc thesis.

Carbonell, Jaime & Jade Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 335–336.

Choobdar, Sarvenaz, Pedro Ribeiro, Sylwia Bugla & Fernando Silva. 2012. Comparison of Co-authorship Networks Across Scientific Fields Using Motifs. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 147–152.

Clarke, James & Mirella Lapata. 2006. Constraint-based Sentence Compression an Integer Programming Approach. In *Proceedings of the COLING/ACL*, 144–151.

Cohn, Trevor & Mirella Lapata. 2008. Sentence Compression Beyond Word Deletion. In *Proceedings of the 22Nd International Conference on Computational Linguistics*, 137–144.

Cohn, Trevor & Mirella Lapata. 2009. Sentence Compression As Tree Transduction. *Journal on Artificial Intelligence Research* 34(1). 637–674.

Cordeiro, João, Gael Dias & Guillaume Cleuziou. 2007a. Biology Based Alignments of Paraphrases for Sentence Compression. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, 177–184.

Cordeiro, João, Gaël Dias & Pavel Brazdil. 2007b. New Functions for Unsupervised Asymmetrical Paraphrase Detection. *Journal of Software* 2(4). 12–23.

Cordeiro, João, Gaël Dias & Pavel Brazdil. 2013. Rule induction for sentence reduction. In Luís Correia, LuísPaulo Reis & José Cascalho (eds.), *Progress in Artificial Intelligence*, vol. 8154, 528–539. Springer.

Costa, Vitor. 2014. *Update Summarization*. Universidade do Porto MSc thesis.

Erkan, Günes & Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *Journal on Artificial Intelligence Research* 22(1). 457–479.

Feldman, Ronen & James Sanger. 2007. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Galley, Michel & Kathleen McKeown. 2007. Lexicalized Markov Grammars for Sentence Compression. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, 180–187.

Gama, João. 2010. *Knowledge Discovery from Data Streams*. Chapman & Hall/CRC.

Grigonytė, Gintarė, João Cordeiro, Gaël Dias, Rumen Moraliyski & Pavel Brazdil. 2010. Paraphrase Alignment for Synonym Evidence Discovery. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 403–411.

Iacobucci, Dawn. 1994. Graphs and Matrices. In *Social Network Analysis*, 92–166. Cambridge University Press.

Jacomy, Alexis. 2013. *sigma js*. http://sigmajs.org.

Knight, Kevin & Daniel Marcu. 2002. Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. *Artificial Intelligence* 139(1). 91–107.

Leal, António. 2011. Some Semantic Aspects of Gerundive Clauses in European Portuguese. *Cahiers Chronos* 22. 85–113.

Leal, José Paulo, Vânia Rodrigues & Ricardo Queirós. 2012. Computing Semantic Relatedness using DBPedia. In *1st Symposium on Languages, Applications and Technologies (SLATE)*, 133–147.

Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of ACL Workshop: Text Summarization Branches Out*, 74–81.

Litvak, Marina, Mark Last, Hen Aizenman, Inbal Gobits & Abraham Kandel. 2011. DegExt - A Language-Independent Graph-Based Keyphrase Extractor. In *Advances in Intelligent Web Mastering*, 121–130.

de Marneffe, Marie-Catherine, Bill MacCartney & Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the IEEE / ACL 2006 Workshop on Spoken Language Technology*, 449–454.

Mihalcea, Rada & Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Conference on Empirical Methods in Natural Language Processing*, 404–411. ACL.

Otterbacher, Jahna, Güneş Erkan & Dragomir R. Radev. 2005. Using Random Walks for Question-focused Sentence Retrieval. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 915–922.

Ouyang, You, Wenjie Li, Sujian Li & Qin Lu. 2011. Applying Regression Models to Query-focused Multi-document Summarization. *Information Processing & Management* 47(2). 227–237.

Pons, Pascal & Matthieu Latapy. 2006. Computing Communities in Large Networks Using Random Walks. *Journal of Graph Algorithms and Applications* 10(2). 191–218.

Price, Simon, Peter A. Flach, Sebastian Spiegler, Christopher Bailey & Nikki Rogers. 2010. SubSift Web Services and Workflows for Profiling and Comparing Scientists and Their Published Works. In *IEEE Sixth International Conference on e-Science*, 182–189.

RStudio, Inc. 2014. *Easy web applications in R.* http://www.rstudio.com/shiny/.

Santos, Diana & Fernando Ribeiro. 2011. Uma incursão pelo universo das publicações em Portugal. *Linguamática* 3(2). 85–98.

Silvano, Purificação. 2010. *Temporal and rhetorical relations: the semantics of sentences with adverbial subordination in european portuguese*: University of Porto PhD dissertation.

Silvano, Purificação. 2012. The rhetorical Relations in complex sentences with quando ('when') in European Portuguese. *Belgian Journal of Linguistics* 26.

Sousa-Silva, Rui, Luis Sarmento, Tim Grant, Aston University, Eugénio Oliveira & Belinda Maia. 2010. Comparing Sentence-Level Features for Authorship Analysis in Portuguese. In *International Conference on Computational Processing of the Portuguese Language (PROPOR 2010)*, vol. 6001, 51–54.

Srinivasan, Ashwin. 2004. The Aleph Manual. Tech. rep. University of Oxford. http://www.comlab.ox.ac.uk/activities/machinelearning/Aleph/.

Toutanova, Kristina, Chris Brockett, Michael Gamon, Jagadeesh Jagarlamudi, Hisami Suzuki & Lucy Vanderwende. 2007. The PYTHY Summarization System: Microsoft Research at DUC 2007. In *Proceedings of DUC*, s/pp.

Trigo, Luís & Pavel Brazdil. 2014. Affinity Analysis between Researchers using Text Mining and Differential Analysis of Graphs. In *ECML/PKDD 2014 PhD session Proceedings*, 169–176.

Valizadeh, Mohammadreza & Pavel Brazdil. 2014. Exploring actor–object relationships for query-focused multi-document summarization. *Soft Computing* 1–13.

Valizadeh, Mohammadreza & Pavel Brazdil. 2015. Density-Based Graph Model Summarization: Attaining better Performance and Efficiency. To be published in IDA Journal.

Ventura, João. 2014. *Automatic Extraction of Concepts from Texts and Applications*: Universidade Nova de Lisboa PhD dissertation.

Ventura, João & Joaquim Silva. 2013. Automatic Extraction of Explicit and Implicit Keywords to Build Document Descriptors. In Luís Correia, Luís Paulo Reis & José Cascalho (eds.), *Progress in Artificial Intelligence*, 492–503. Springer.

Wan, Xiaojun, Jianwu Yang & Jianguo Xiao. 2006. Using Cross-Document Random Walks for Topic-Focused Multi-Document. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 1012–1018.

Wasserman, Stanley & Katherine Faust. 1994. *Social network analysis: Methods and Applications*. Cambridge University Press.

Wei, Furu, Wenjie Li, Qin Lu & Yanxiang He. 2008. A Cluster-Sensitive Graph Model for Query-Oriented Multi-document Summarization. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven & Ryen W. White (eds.), *Advances in Information Retrieval*, vol. 4956, 446–453. Springer.

CONTACTS

Pavel Brazdil
LIAAD-INESC Tec; FEP, Univ. of Porto
pbrazdil@inescporto.pt

Luís Trigo
LIAAD-INESC Tec
lptrigo@inescporto.pt

João Cordeiro
LIAAD-INESC Tec; Univ. of Beira Interior
jpaulo@di.ubi.pt

Rui Sarmento
LIAAD-INESC Tec
rui_sarmento@hotmail.com

Mohammadreza Valizadeh
LIAAD-INESC Tec; Univ. of Ilam
valizadehmr@gmail.com