SEPIR: a SEmantic and Personalised Information Retrieval Tool for the Public Administration based on Distributional Semantics

Pierpaolo Basile and Annalina Caputo

Department of Computer Science, Univeristy of Bari Aldo Moro,
Via E. Orabona, 4, 70125 Bari, Italy
E-mail: {pierpaolo.basile,annalina.caputo}@uniba.it

Marco Di Ciano and Gaetano Grasso

InnovaPuglia S.p.A.,
c/o Tecnopolis, str. Prov. per Casamassima km. 3, 70010 Valenzano,
Bari, Italy
E-mail: {m.diciano,g.grasso@innova.puglia.it}@innova.puglia.it

Gaetano Rossiello and Giovanni Semeraro

Department of Computer Science, Univeristy of Bari Aldo Moro,
Via E. Orabona, 4, 70125 Bari, Italy
E-mail: {gaetano.rossiello,giovanni.semeraro}@uniba.it

Abstract: This paper introduces a semantic and personalised information retrieval (SEPIR) tool for the public administration of Apulia Region. SEPIR, through semantic search and visualisation tools, enables the analysis of a large amount of unstructured data and the intelligent access to information. At the core of these functionalities is an NLP pipeline responsible for the WordSpace building and the key-phrase extraction. The WordSpace is the key component of the semantic search and personalisation algorithm. Moreover, key-phrases enrich the document representation of the retrieval system and are on the basis of the bubble charts, which provide a quick overview of the main concepts involved in a document collection.We show some of the key features of SEPIR in a use case where the personalisation technique re-ranks the set of relevant documents on the basis of the userâĽ™s past queries and the visualisation tools provide the users with useful information about the analysed collection.

Keywords: semantic search; personalised information retrieval; natural language processing; distributional semantics; intelligent information access.

Reference to this paper should be made as follows: Basile, P., Caputo, A., Di Ciano, M., Grasso, G., Rossiello, G. and Semeraro, G. (2017) 'SEPIR: a semantic

*Accepted author manuscript.

*Electronic version of an article published as International Journal of Electronic Governance, Vol.9, No.1/2, pp.132 - 155, DOI: 10.1504/IJEG.2017.10005481
http://www.inderscience.com/offer.php?id=84654

and personalised information retrieval tool for the public administration based on distributional semantics', *Int. J. Electronic Governance*, Vol. 9, No. 1-2, pp.132 - 155.

Biographical notes: Pierpaolo Basile received his PhD in Computer Science at the University of Bari Aldo Moro. He is an Assistant Professor at the Department of Computer Science, University of Bari. His research interests include Statistical methods for Natural Language Processing, Information Retrieval and content-based Recommender Systems. He is author of several research studies published at national and international journals, conference and workshops proceedings.

Annalina Caputo is a research fellow in the ADAPT centre at Trinity College Dublin in the domain of Personalised Information Retrieval. Her areas of interest are Natural Language Processing and Information Retrieval. Before joining ADAPT, she worked as a research fellow at the University of Bari Aldo Moro where she focused on both entity linking and developing distributional semantic models for intelligent information access. She obtained her Ph.D. in Computer Science defending the thesis titled "Semantics and Information Retrieval: Models, Techniques and Applications".

Marco Di Ciano received the Laurea cum laude in Electronic Engineering at Politecnico di Bari in 1993. In 1998, he got his PhD working also as free researcher at the Katholieke Universiteit Leuven Belgium. As researcher and microelectronics engineer, he cooperated with the Italian National Council of Research, obtained a research contract at IMEC-Leuven in Belgium and got a permanent position as group leader at Philips, Zurich, Switzerland. In 2000, he joined the Scientific Park of Tecnopolis CSATA in Bari. Since 2009, he works for InnovaPuglia S.p.A., the in-house company of Regione Puglia where he leads the Research and Innovation Service.

Gaetano Grasso received the PhD in Medicinal Chemistry at the Faculty of Pharmacy, University of Bari, he joined in 1988 the Microelectronics Laboratory at Tecnopolis Science Park in Bari where he applied his chemistry expertise in materials research for electronics devices in Space application. He developed a specialized professional skill in science and technology management at the European Innovation Relay Center. From 2012, he is Project Leader for InnovaPuglia S.p.A. at the Research and Development Office, leading regional industrial funding programmes for ICT application. In 2014, he started the Project Management of the running ICT Apulian Living Labs Programme.

Gaetano Rossiello received his M.Sc. degree in Computer Science at the University of Bari Aldo Moro with full marks and honors. He is a PhD student at Department of Computer Science, University of Bari. He is a member of Semantic Web Access and Personalization (SWAP) research group. His research interests include Probabilistic and Deep Learning methods for Natural Language Processing, Information Retrieval and Recommender Systems.

Giovanni Semeraro received the M.Sc. degree in computer science from the University of Bari Aldo Moro, Italy. He is a Full Professor of Computer Science at the University of Bari Aldo Moro, where he teaches "Intelligent information access", "Natural language processing", and "Programming languages", and he leads the Semantic Web Access and Personalization (SWAP) research group. He has been a Visiting Scientist with the Department of Information and Computer Science, University of California at Irvine, in 1993. From 1989 to 1991, he was a researcher at Tecnopolis CSATA Novus Ortus, Bari, Italy. His research interests include AI; recommender systems; intelligent information mining, retrieval, and filtering; semantic and social computing; machine learning; natural language processing; the semantic web.

1 Introduction

The 2013 publication of the European Commission vision for public services (`https://ec.europa.eu/digital-single-market/en/news/vision-public-services`) states that the evolution of society requires public administrations to tackle many new challenges with respect to public e-services. This is also due to economic and budgetary pressures which force governments to be even more efficient in future investments by leveraging research and innovation strategies for local socio-economic developments.

In the last programming period of the European Regional Development Fund (ERDF), the Department of Economic Development of Apulia Region has devoted specific attention to support and finance digital e-services innovation in order to pave the way to a more sustainable, inclusive and intelligent growth of the territories envisioned in the Smart Specialization Strategy SmartPuglia2020. A huge amount of information coming from granted projects for digital innovation, specific support actions devoted to SMEs (Small and Medium-sized Enterprises) industrial research and technological framework programmes for Public Private Partnership are often stored in terabytes of digital documents. Design and monitoring the innovation strategy by benchmarking evidences with existing outcomes as well as with already collected results coming from implemented policy actions require a permanent and updated knowledge of local territories and capabilities. Some of these evidences (research contents, enabling technologies, network relationship analysis, etc.) could be retraced starting from appropriate studies and detailed analysis of implemented projects documented in digital content databases. Reference mapping of territorial evidences can therefore benefit from semi-automatic mechanisms of keyphrases extraction, concept identification, textual similarity evaluation, implemented through iterative and multiple process analyses on specific digital document collections. The future of government is less and less in the hands of governments alone.

End users as academia, entrepreneurs, research centers and local associated territories that make their activities, ideas and needs monitored and considered by the government leaders can become more and more effective with the adoption of new digital technologies not only to produce, but also to understand the ability and willingness of regional stakeholders to address public concerns. In this scenario, e-Government technologies play an important role since they provide citizens and entrepreneurs with convenient access to government information and services that help to improve the quality of services themselves and enhance the decision making involved in the governance processes [Fan02]. Then, Natural Language Processing (NLP), semantic technologies, machine learning techniques, information retrieval and recommender systems are some of the potential methodologies offered by computer science, artificial intelligence, and computational linguistics able to outsmart the digital deluge by supporting interactions of humans with digital data in e-Government processes. The main motivation behind the use of both NLP and semantic technologies is to dig deep into the textual content of documents in order to discover interesting and non-trivial pieces of information. For example, semantic analysis of technical descriptions of funded projects can reveal new knowledge not available in structured form in databases. The exploited technologies, the semantic relatedness between terms, the

semantic similarity between documents can be exploited in order to discover similar projects. Moreover, the semantic analysis of documents enables the development of more effective information retrieval and personalization services.

This paper introduces SEPIR, a SEmantic and Personalised Information Retrieval tool able to index, analyse and search heterogeneous document collections in order to extract knowledge from the textual content through natural language processing, distributional semantics, personalization and data visualization techniques. A first prototype of SEPIR has been developed for the public administration of Apulia Region in order to extract knowledge and strategical information from documents that will support the decision making and the strategic planning in e-Government activities, in particular for e-Procurement purposes and funded projects analysis. Apulia Region requires innovative tools for the textual analysis of documents coming from call for tenders or patents to perform its revision and control activities for the public-private partnership programme. Moreover, the Apulia Region is interested in analysing the impact of funded projects in terms of technologies successfully adopted by the region. Due to the large amount of this information collected from different sources, the semantic search and personalization capabilities of SEPIR are the workhorse that provides users with effective tools for intelligent information access.

The key features of SEPIR are summarized as follows:

Indexing: A document storage module that relies on an inverted index data structure. Since the documents may come from heterogeneous sources, the tool supplies features to organize documents in different collections.

Semantic Search: In addition to the classical search, based on the exact matching between terms in query and documents, the tool offers a semantic search functionality that retrieves documents according to their semantic relatedness with the query terms. For this purpose a distributional semantic space from any document collection is build by using the Random Indexing technique [Sah05].

Personalized Search: The tool implements a novel personalization capability combining explicit relevance feedback and distributional semantics methods. A semantic user profile is inferred from previous searches and is used to improve the effectiveness of document retrieval across the different document collections.

Data Visualization: A component that supplies several graphical tools for visualizing the main concepts in a collection as well as different types of correlations between documents through the semantic space.

RESTful API: A middleware layer that exposes the services using the REST protocol. This allows a remote access to platform services through RESTful APIs in order to facilitate the system integration across the departments of the public administration.

The remainder of the paper is organized as follows: Section 2 describes the methodology behind the personalisation process at the core of SEPIR and the system architecture. Section 3 sets the scene of some use cases to test the potentialities of SEPIR, while Section 4 proposes a preliminary quantitative evaluation of the personalization algorithm. Section 5 reviews the state-of-the-art in semantics and personalisation in information access for e-Government, followed by conclusions about the proposed system.

2 Methodology

In this section we provide details about SEPIR and the personalisation methodology adopted in the searching process. Before describing the personalization methodology in Section 2.2, we provide details about the system architecture and its components.

2.1 System Architecture

Figure 1 shows the main components of our system. All the functionalities are exposed as services through a REST API. The front-end has been developed as a Web application that relies on the REST API. The Web application provides an easy access to all the functionalities of SEPIR. In particular, the Web application has been customized for the needs of Apulia Region, in order to support multiple collections of documents and enables the search and analysis across one or more of them.

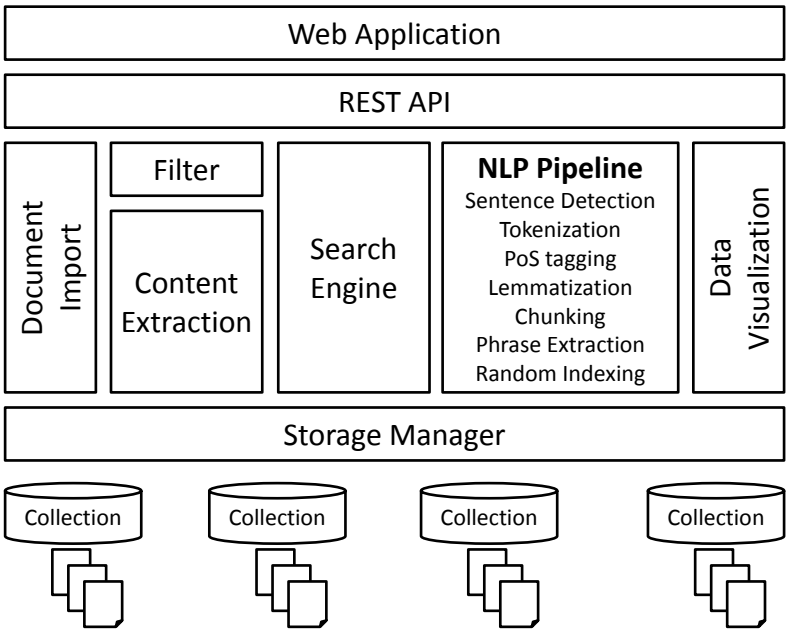

Figure 1 A schema of SEPIR architecture.

A core component of the architecture is the *Storage Manager*, that deals with all the I/O operations on the collections of documents. The *Storage Manager* supplies: 1) an abstract representation level for both documents and collections of documents; 2) access to the indexing and search capabilities; 3) a communication interface with both the *Content Extraction* and the components of the *NLP pipeline*. The current implementation of the *Storage Manager* stores documents on the file system. Documents can be grouped into collections, where each collection has a name and a default language and is composed of

a set of documents, an index, a *WordSpace*, and a file containing the information extracted by the NLP pipeline. Moreover, a collection can be public or private. A public collection is accessible to all registered users, while a private collection is accessible only to the user who created it. In a public collection, it is possible to assign different privileges to each user: administrator, only reader, reader and writer.

Documents can be added to any collection. In particular, the system is able to import documents (*Document Import* component) from a directory, a single file or a CSV. The import from a CSV allows to split a document in several user defined sections. Example of section are: *title*, *abstract*, and *content*. During the import, the user can select the fields of the CSV she/he is interested in. When the user imports documents from a directory or a single file, the system is able to recognize the file format and to extract the textual content. The extraction process is implemented using Apache Tika (`https://tika.apache.org/`). The Apache Tika toolkit is able to detect and extract both metadata and text from several file types (DOC, PDF, PPT, and XLS). All these file types can be parsed through a single interface, making Tika useful for content analysis. During the import process, our system can apply customised filters (*Filter* component) able to identify the different sections of a document, in a manner similar to what happens for CSV. The filters can be defined by users exploiting regular expressions to identify the beginning and the end of a document section. In such a way, it is possible to filter out redundant sections or build up a collection containing only specific information.

The *Search Engine* provides support for the indexing and the retrieval of documents. Each collection can be indexed using a classical Vector Space Model [SWY75] implemented by Apache Lucene (`https://lucene.apache.org/core/`). Users can search over the collection using keywords and boolean operators. Moreover, this component supports also a *semantic* search performed through a distributional space where related concepts are represented as near points in the space. The distributional space, which is the key component of the semantic search, is built by the *NLP pipeline* through Random Indexing (RI). More details are provided in Subsection 2.1.2. Both classical and semantic search take into account the user profile, which is built by RI technique. Details of the personalisation process are given in Subsection 2.2.

The *Data Visualisation* component exploits graphical tools for the semantic analysis of documents in a collection.

Finally, the core of our platform is the Natural Language Processing component (*NLP pipeline*), which feeds both the *Search Engine* and the *Data Visualization* components. The pipeline is able to perform several text processing steps for English and Italian:

Sentence Detection splits a text in sentences, by exploiting punctuation characters that mark the end of a sentence.

Tokenization splits the text into tokens. Each token is a word.

Part-of-Speech (POS) tagging identifies the grammatical role of each word: noun, verbs, adjective, adverb, punctuation, preposition, and so on.

Lemmatization provides the lemma for each word. The lemma is the basic form of a word, for example the singular form of a noun or the infinitive form of a verb, as shown at the beginning of a dictionary entry.

Chunking divides a text in syntactically correlated parts of words, like noun or verb groups, but it specifies neither their internal structure nor their role in the main sentence.

Phrase Extraction is able to find n-grams (sequence of words) that identify a single concept. Examples of n-grams are:: *Information Retrieval*, *Document Management*, *Public Administration*. Section 2.1.1 explains in details this module.

Random Indexing constructs a *WordSpace* by analysing a collection of documents. A *WordSpace* is a geometrical space in which words are represented as points. If two words are close in the *WordSpace* they are semantically related. Section 2.1.2 gives further details about this module.

For example, given the following piece of text extracted from a patent description: *"A power load management system for regulating power demand from a distribution panel of a residence or building is disclosed. Load control switches placed inline between circuit breakers of the distribution panel and the loads they control, such as a water heater, pump, AC unit."*, the NLP pipeline is able to identify two sentences. For each sentence, the list of tokens and lemmas are extracted. The chunking module is able to identify noun phrases such as *distribution panel*, or verb phrases like *is disclosed*. The phrase extraction is able to automatically identify relevant concepts such as *power load management system* or *water heater*.

2.1.1 Phrase Extraction

The phrase extraction component implements two methods to extract key-phrases. The first method is completely unsupervised and is based on the idea that words that occur frequently together, and infrequently in other contexts, are good candidates for a phrase.

We use a simple data-driven approach, which builds up phrases based on the uni-gram and bi-grams counts, following the same approach proposed in [MSC+13]. We chose this method for two reasons: 1) it is completely unsupervised and does not require any external resource such as dictionaries or gazetteers; 2) it is simple because it is based only on word frequencies. All bi-grams are scored using Equation 1, where w_i and w_j are two words that occur in the collection under analysis, and $count()$ is the function that returns the number of occurrences in the collection for uni-gram $count(w_i)$ and bi-grams $count(w_i, w_j)$.

$$score(w_i, w_j) = \frac{count(w_i, w_j) - minCount}{count(w_i) \times count(w_j)} \tag{1}$$

The $minCount$ factor prevents the formation of phrases consisting of very infrequent words. Bi-grams whose score is above a chosen threshold are used as phrases. Both the $minCount$ and the threshold can be defined by the user. Through the method is able to extract only bi-grams, it is possible to recognize phrases composed by more than two terms by concatenating two or more bi-grams. For example, if the algorithm discovers the two bi-grams *information retrieval* and *retrieval system*, the tri-gram *information retrieval system* can be still recognized.

The second approach exploits a Finite State Automaton (FSA) able to recognize sequences of words that are part of Wikipedia categories or a generic list of concepts. This approach is useful when a list of predefined concepts is available. We extract all the Wikipedia category labels from both the Italian and English versions of Wikipedia. We consider only categories with more than one word and up to six words. These categories are used to build the FSA; in this way the FSA is able to recognize sequences of characters that are labels of Wikipedia categories. The idea behind this approach is that Wikipedia categories can express concepts consisting of more than one word.

The phrases extracted in this step can be used to tokenize the text. This operation can affect both the indexing and the *WordSpace* creation.

2.1.2 Random Indexing

The objective of Random Indexing (RI) [Sah05] is to represent words as points in a *WordSpace*, a vector space where two words are semantically related if they are represented by close points. For example, we expect that the vector for the word *dog* is close to the vector for the word *cat*. In other words, semantic relatedness of words is represented by closeness of their vectors in the *WordSpace*. RI has the advantage of being very simple, since it is based on an incremental approach. In addition it does not require external knowledge or resources, such as dictionaries, thesauri or ontologies. In this way, RI is completely unsupervised and language independent since the only required pre-processing operation is the tokenization. The *WordSpace* is built by taking into account word co-occurrences according to the distributional hypothesis [Har68] which states that words sharing the same linguistic contexts are related in meaning. In our case the linguistic context is defined as the words that co-occur.

The idea behind RI has its origin in Kanerva work about Sparse Distributed Memory [Kan88]. RI assigns a random vector to each context unit, represented by a word in our case. The vector is generated as a high-dimensional random vector with a high number of zero elements and a few number of elements equal to 1 or -1 randomly distributed over the vector dimensions. Vectors built using this approach generate a nearly orthogonal space since the probability of the cosine similarity between any two random vectors being near to 0 is very high. During the incremental step, a vector is assigned to a word as the sum of the random vectors representing the context in which the word is observed. In our case the target element is a word and the contexts are the co-occurring words that we observe by analysing all the documents belonging to a collection.

Finally, we compute the cosine similarity between the vector representations of word pairs in order to compute their relatedness.

Formally, the mathematical insight behind the RI is the projection of a high-dimensional space onto a lower dimensional one using a random matrix (Figure 2); this kind of projection does not compromise distance metrics [DG99].

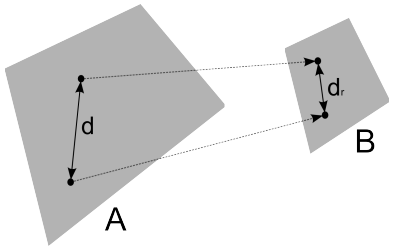

Figure 2 Random Projection.

Formally, given a $n \times m$ matrix A and an $m \times k$ matrix R, which contains random vectors, we define a new $n \times k$ matrix B as follows:

$$A^{n,m} \cdot R^{m,k} = B^{n,k} \quad k << m \tag{2}$$

The new matrix B has the property of preserving the distance between points, that is to say, if the distance between any two points in A is d, then the distance d_r between the corresponding points in B will satisfy the property that $d_r \approx c \times d$. A proof of that is reported in the Johnson-Lindenstrauss lemma [DG99].

Specifically, RI creates the *WordSpace* in two steps:

1. A random vector is assigned to each word in the collection. This vector is sparse, high-dimensional and ternary, which means that its elements can take values in {-1, 0, 1}. A random vector contains a small number of randomly distributed non-zero elements, and the structure of this vector follows the hypothesis behind the concept of Random Projection;

2. Context vectors are accumulated by analyzing co-occurring words. Specifically, the semantic vector for any word is computed as the sum of the random vectors for words that co-occur with the analyzed word.

Formally, given a collection C of n documents, and a vocabulary V of m words extracted from C, we perform two steps: 1) assign a random vector r to each word w in V; 2) compute a semantic vector sv_i for each word w_i as the sum of all random vectors assigned to words co-occurring with w_i. The context is the set of c words that precede and follow w_i. The second step is defined by the following equation:

$$sv_i = \sum_{d \in C} \sum_{\substack{-c<j<+c \\ j \neq i}} r_j \tag{3}$$

After these two steps, we obtain a set of semantic vectors assigned to each word in V representing a *WordSpace*.

For example, let us consider the following sentence: *"The quick brown fox jumps over the lazy dog"*. In the first step we assign a random vector to each term as follows:

$$\begin{aligned}
r_{quick} &= (-1, 0, 0, -1, 0, 0, 0, 0, 0, 0) \\
r_{brown} &= (0, 0, 0, -1, 0, 0, 0, 1, 0, 0) \\
r_{fox} &= (0, 0, 0, 0, -1, 0, 0, 0, 1, 0) \\
r_{jumps} &= (0, 1, 0, 0, 0, -1, 0, 0, 0, 0) \\
r_{over} &= (-1, 0, 0, 0, 0, 0, 0, 0, 0, 1) \\
r_{lazy} &= (0, 0, -1, 1, 0, 0, 0, 0, 0, 0) \\
r_{dog} &= (0, 0, 0, 1, 0, 0, 0, 0, 1, 0)
\end{aligned}$$

In the second step we build a semantic vector for each term by accumulating random vectors of its co-occurring words. For example, fixing $c = 2$ the semantic vector for the word *fox* is the sum of the random vectors *quick, brown, jumps, over*. Summing these vectors, the semantic vector for *fox* results in $(-2, 1, 0, -2, 0, -1, 0, 1, 0, 1)$. This operation is repeated for all the sentences in the collection and for all the words in V. In this example, we used very small vectors, but in a real scenario the vector dimension ranges from hundreds to thousands of elements. In SEPIR for each collection the user can select: the vector dimension d, the number of non-zero elements (seed), and the dimension of V.

The *WordSpace* can be easily used for building a vector for each document in the collection. In particular, a semantic vector sv_{d_j} for a document j can be computed as the sum of semantic vectors sv_i for the terms occurring in d_j. During the document vector construction each semantic vector can be weighted by using the inverse document frequency of the term w_i. This approach gives a boost to terms that are more relevant with respect to the whole collection.

The representation of both terms and documents in the same *WordSpace* allows the implementation of several semantic information access tools that we developed in SEPIR. In particular, we make available two functionalities:

- **Document Similarity**: given a document vector sv_{d_j} we rank all other document vectors with respect to their cosine similarity with sv_{d_j}. In this way it is possible to find the most similar documents with respect to d_j;
- **Semantic Search**: given a query q composed of k terms, we build a query vector $\vec{q}$ in the *WordSpace* as the vector sum of the k terms belonging to q. Then we can rank all the document vectors with respect to their cosine similarity with respect to q.

The methodologies adopted here to implement document similarity and semantic search are at the basis of the personalization techniques implemented in our system.

2.2 Personalisation

Our search engine exploits a personalisation algorithm in order to take into account the user profile during the retrieval of relevant documents. The profile consists of the user past queries. The queries provide a contextual information that is taken into account in the ranking algorithm. We decided to adopt a content-based strategy. The idea is to give more importance to documents that are similar to past queries since they represent the user interests. The similarity is computed by exploiting the semantic relatedness between documents and the user profile in the *WordSpace*.

Given a ranked list of documents R, we want to re-rank this list according to the previous user queries. In our system the rank R can be provided indifferently from the classical search engine or the semantic one. The personalization strategy can be applied in both cases. Moreover, we want to weigh the contribution of each past query to the user profile according to the time the query was performed. The idea is that a query executed a lot of time ago is less relevant than a more recent query.

Let P_i be the profile of a user u_i. The profile contains information about queries made by the user u_i. For each query we store a query vector and its time stamp. The query vector is built according to the method proposed in Section 2.1.2 by exploiting the vector sum. Then we assign a vector $\vec{P_i}$ to each user profile. This vector is computed as the weighted sum of query vectors belonging to P_i. The weight is assigned according to Equation 4

$$w(q_i, P_i) = e^{-\frac{t}{\gamma}} \tag{4}$$

where t is the difference in days between the current date and the query time stamp, and γ is a factor that determines how rapidly the relevance of q_i decays. We set this factor to 10, this means that after 30 days (one month) the relevance of q_i is near to 0. This approach aims to reflect the time factor in the user profile by giving more importance to recent queries and it is inspired to time-adaptive collaborative filtering algorithms [DL05].

Finally, we re-rank R according to the similarity of each document vector in R with respect to the user profile vector $\vec{P}_i$. In particular, for each document d_r in R we linearly combine the score of the document retrieved by the search engine with the cosine similarity between sv_{d_j} and $\vec{P}_i$.

2.3 Data Visualization

SEPIR provides data visualization tools to easily visualize the content of a document collection. Specifically, SEPIR enables the visualisation of phrases extracted from the NLP pipeline. Then, it is possible to quickly grasp the main concepts that occur in the document collection. Moreover, a tool allows to visualise the semantic similarity between documents; in that way, it is possible to detect cluster of similar documents. This analysis is particularly helpful to relate documents that belong to different collections in order to discover similar contents through different domains. In particular, the data visualization tool is able to provide three kinds of charts:

1. A bubble chart that shows the phrases extracted from the collection by exploiting the FSA. Each bubble represents a phrase and its color and diameter depends on the occurrences of the phrase in the collection.
2. A bubble chart of phrases extracted by the unsupervised approach.
3. A correlation matrix between documents in the collection. In this case, the color indicates the level of the correlation. The correlation is computed by exploiting the document similarity in the *WordSpace*.

3 Use Cases

A first prototype of SEPIR has been developed for the public administration of Apulia Region in order to extract knowledge and strategical information from documents that support the decision making and the strategic planning in e-Government activities.

We propose two use cases that show how the synergistic use of semantic, personalisation and visualisation tools provided by SEPIR helps in making sense of the wealth of information managed in public administration. The former carries out an analysis of patent data related to the pre-commercial procurement procedures, the latter analyses several batches of projects funded by Apulia Region in order to understand cases of success in the financed technologies.

3.1 Patent Data Analysis

The patent data analysis is related to pre-commercial procurement procedures that require a preliminary step of market consultation in order to assess the technology-market gap through the patent screening in the field of interest. Indeed, a higher number of patents could imply a spread use of the technology by the market, hence this can be an indicator for the applicability of pre-commercial procurement procedures. However, problems like the lack of a domain knowledge, vocabulary drift, polysemy and synonymy can hamper the judgement of the real extent of applicability of a technology. This kind of analysis requires innovative tools for the textual analysis of documents coming from call for tenders

or patents to perform screening and control activities for assessing the distance between the technologies and the market in a program of pre-commercial procurement. In particular, due to the large amount of information collected from several sources, the semantic search and personalization capabilities of SEPIR provide effective tools for intelligent information access. This use case aims to show the effectiveness of the semantic tools provided by our system for alleviating the burden of analysing huge amounts of patents to estimate the coverage of a given technology. The use case shows how SEPIR can be used to obtain two indicators of the technology-market gap: 1) the number of patents that exploit a technology and 2) the widespread of a given technology across different domains. Then, this study focuses on the following aspects:

1. Effectiveness at retrieving relevant patents. The more accurate the system, the more precise the assessment of the extent of a given technology. In this use case, we use accuracy as a measure of effectiveness for the relevance of documents to some given topics.

2. Relevance of detected technologies. A measure of relevance of a technology can be its adoption in different domains (i.e. topics). We analyse the correlations, in terms of semantic similarity, between patents that exploits the same technology but in different topic collections.

Moreover, we show as SEPIR makes easier the retrieval and exploration of the collection content through its personalisation and visualisation tools.

The analysis starts defining an area of interest (topic) and a set of technologies. Relying on these information, we perform several queries to the European Patent Office (EPO) (`https://www.epo.org/searching/free/ops.html`) in order to retrieve relevant patents. In particular, we perform two kinds of query using the operators *any* (OR) and *all* (AND). For each pair of topic and technology, we perform a query and retrieve the top 1,000 documents. For each topic, we group and merge documents in order to obtain a unique collection of documents.

The use case involves three topics: *sludge reduction*, *water leak* and *adaptive water management*. We collect for each topic respectively 2,593, 2,083 and 1,634 documents.

Each collection is processed by exploiting the semantic tools of the SEPIR framework. For each collection, we build the semantic index and the *WordSpace* by using the RI method. We run the *NLP pipeline* with default values: 1) phrase threshold and $minCount$ set to 100 and 5, respectively; 2) RI vector dimension and number of non zero elements set to 200 and 2; 3) the words taken into account in RI were the 50.000 most frequent words.

The first step concerns the extraction of related concepts for each collection. This step is performed by extracting phrases and related words for each technology. Related words are selected by computing the semantic relatedness in the *WordSpace* with respect to each topic. Table 3.1 reports the top ten most relevant concepts for each topic.

The second step involves the semantic search engine. We build a unique collection by merging the three collections built in the previous step. Two experts query the search engine using the related concepts extracted in the first step and for each query retrieve the top five documents. We compute the accuracy of the system as the ratio between the number of documents judged as relevant by the experts and the total number of documents retrieved during the evaluation. We measure an encouraging accuracy of 0.73. This is a preliminary result, since the evaluation involves only two experts, but we plan to extend the evaluation as future work.

Sludge Reduction	**Water leak**	**Adaptive water managment**
Water treatment	Water treatment	Network access
Swollen super absorbent	Heat transfer	Water treatment
OFDM	Viral trapper	Viral trapper
Mololithic selector	Flame retardant	Flame retardant
Powder boric	Soilless lawn	Urinal flusher
Zerovalent	Asphalt pavement	Bulletproof vest
Flame retardant	Sol gel	Coal slack
Seamless titanium	Polycrystalline ingot	Access control
Oxering receptor	Blank stencil	Clock signal
Primer kit	Nuclear power	Sunlight reflection

Table 1 The top ten most relevant concepts for each topic.

Another crucial aspect in the patent analysis task is the identification of technologies that are used in different domains (i.e. topics) since this represent a further indicator of the relevance and impact of a technology. For this purpose we build a correlation matrix by exploiting the semantic similarity between documents computed in the *WordSpace*. For each document, we consider only the top ten most similar documents. We restrict the analysis only to those documents that have at least one technology in common and belong to different topics. For these documents, we analyse the level of correlation and we discover that the 54.4% of them are highly correlated (similarity between 0.6 and 1.0), while the 26% of the documents are slightly correlated (similarity between 0.6 and 0.4). This quantitative analysis proves that the extracted technologies are also the most promising since they are correlated to different areas of interest.

We decide to set up a case study to show the potentialities of both the personalized search engine and the visualization tools. The case study involves the indexing of patents belonging to the three different topics used in the previous evaluation and a set of example queries with and without profile. Moreover, we show the output of the three visualisation charts explained in Subsection 2.3.

We simulated a user that performs the query *filter* with the intent of retrieving documents about filtering mechanisms for water. Table 2 reports the top five relevant documents retrieved with the two methods available in SEPIR: 1) the classical Vector Space Model developed in Lucene and 2) the Semantic Search implemented through Random Indexing.

Then, within a user session we addressed the following queries: *reverse osmosis*, *ion exchange membranes*, *membrane*, *water waste treatment*. All these queries regard methods for filtering waste water. Then, we performed again the *filter* query in order to assess variations in the rank of documents. Table 3 reports the top five relevant documents retrieved with the two methods after the user session has started. Both ranks changed according to the user profile by promoting the document $ID5$ that proposes a filter for electro-osmosed sludge; but more interestingly, the rank of the Semantic Search promoted also two new relevant documents ($ID7$ and $ID8$), which are also about sludge filtering.

Finally, we show some screen shots of the visualisation tools applied to this collection. Figure 3 shows the bubble chart obtained by the FSA with the Wikipedia categories on the whole patent collection. The matrix in Figure 4 shows the correlation between the top 15 documents that were retrieved by the VSM and the Semantic Search engines. We limited the number of documents for optimising the visualization. From the matrix, it can be noticed that documents $ID5$, $ID7$ and $ID8$ are all strongly correlated. Other groups of correlated

VSM		***Semantic Search***	
ID	*Content*	*ID*	*Content*
ID1	Filter assembly for filter as you pour filtration	ID1	Filter assembly for filter as you pour filtration
ID2	Filter for distilling water for industrial battery	ID6	Filter media for filter as you pour filtration
ID3	Filter cleaning apparatus of water purifier and mehtod thereof	ID5	Membrane filter plates used for electro-osmosed sludg
ID4	System for filtering vocs with water circulation system	ID7	Membrane filter plates used for electro-osmosed sludge
ID5	Membrane filter plates used for electro-osmosed sludge	ID8	Sludge filter-pressing method of membrane filter press

Table 2 Top five relevant documents for the query *filter* with a classical search engine based on VSM and with the semantic search engine.

VSM		***Semantic Search***	
ID	*Content*	*ID*	*Content*
ID5	Membrane filter plates used for electro-osmosed sludge	ID5	Membrane filter plates used for electro-osmosed sludge
ID1	Filter assembly for filter as you pour filtration	ID7	Membrane filter plates used for electro-osmosed sludge
ID3	Filter cleaning apparatus of water purifier and mehtod thereof	ID1	Filter assembly for filter as you pour filtration
ID2	Filter for distilling water for industrial battery	ID6	Filter media for filter as you pour filtration
ID4	System for filtering vocs with water circulation system	ID8	Sludge filter-pressing method of membrane filter press

Table 3 Top five relevant documents for the query *filter* with a classical search engine based on VSM and with the semantic search engine within a user session where the queries *reverse osmosis*, *ion exchange membranes*, *membrane*, *water waste treatment* were performed.

documents are: $ID9$, $ID13$ and $ID15$, and $ID1$, $ID6$ and $ID12$. On the same set of documents we extracted phrases with the unsupervised approach, the chart is showed in Figure 5.

Analysing the two bubble charts, we can note that some not significant phrases are extracted, for example "used for" and "can be". Generally, these phrases are frequent grammatical expressions that involve function words. We cannot simply remove phrases that contain functional words since these can also occur in some relevant technical expressions. We plan to improve our methodology in order to mitigate this issue. However, bubble charts obtained by the FSA provide more significant results since they are built using a predefined list of phrases.

3.2 Financed Projects Analysis

The second analysis involves three collections: two of them are a set of technical reports about financed projects by the Apulia Region while the other one contains descriptions of needs provided by several stakeholders. In this case all the documents are in Italian.

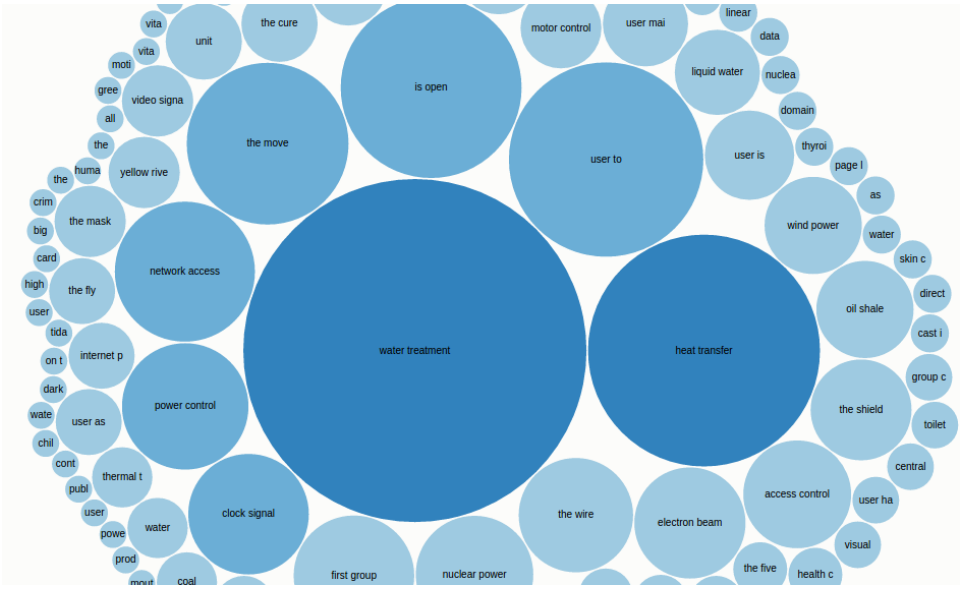

Figure 3 An example of bubble chart built on Wikipedia categories of the whole collection.

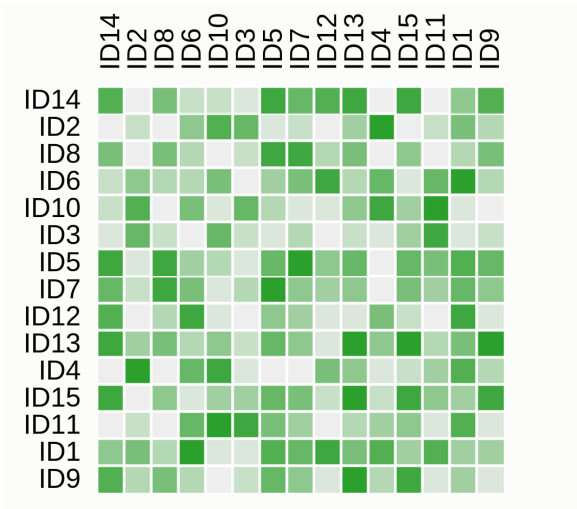

Figure 4 An example of correlation matrix between documents.

The Apulia Region is interested in discovering which technologies were financed in research projects in the past and were subsequently exploited by other financed projects. The idea is to evaluate the funding effectiveness in terms of the impact of the research activities on the market. Details about the involved collections are as follows:

SP 45 technical reports about research projects financed by a call for tenders concerning strategic projects. Typically this call is directed to research centres and universities. The call was published in the 2007;

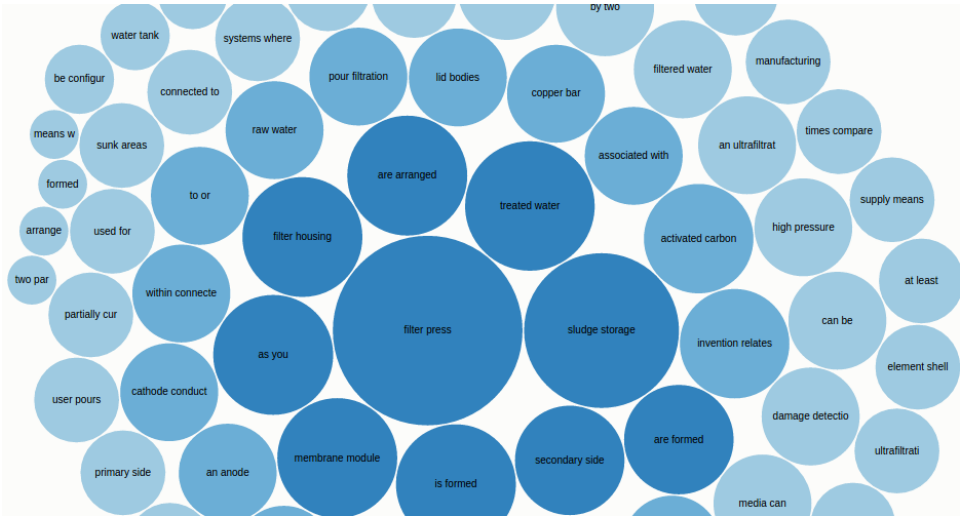

Figure 5 An example of bubble chart built using the unsupervised approach.

NE 473 descriptions of needs provided by several stakeholders collected before the publication of the Living Labs call;

LL 74 technical reports about research projects financed by the Living Labs call published by the Apulia Region in the 2013.

It is important to underline that the Living Labs proposals are written taking into account the needs provided by the stakeholders. The analysis involves three steps: 1) extraction of the most relevant technologies from the SP collection; 2) identification of LL projects that are correlated to the needs reported in the NE collection; 3) identification of common technologies between LL projects resulting from step 2 and the technologies extracted in the step 1.

In the first step, we extract the technologies from the SP collection by exploiting the bubble chart built on Wikipedia categories. The regional expert has chosen to consider only the first twelve most relevant concepts reported in Table 3.2.

Data Mining	X-Ray	Mass Spectrometry
Thermal treatments	Nuclear Magnetic Resonance	Solar Energy
Augmented Reality	Cloud Computing	Optic Fiber
Waste Management	Energy Saving	Electric Vehicles

Table 4 The twelve most relevant technologies extracted from the SP collection.

In the second step, we use the correlation matrix between documents belonging to the NE and LL collections, we consider all the pairs of documents that have a correlation value equal or greater than 0.7. We chose to restrict our analysis to only those projects that are highly correlated with stakeholder needs because we are interested in technologies that are valued by the market. Finally, we identify the technologies extracted in the first step that are most frequent in the LL projects deriving from the step 2. These technologies are: *Data mining*, *Cloud computing*, *Augmented Reality*, *Mass Spectrometry*, and *Solar Energy*. In

particular, *Data mining* and *Cloud computing* are the most frequent in both SP and LL projects. This outcome suggests that an investment in these technologies turned out in a successful exploitation by the market. The presence of these technologies in both SP and LL projects suggests that an initial investment in research projects of strategical importance (SP) turned into a technology need by the market that has decided to further invest in them (LL).

4 Evaluation

In this section we propose a quantitative evaluation with the aim of measuring the impact of personalisation on the document ranking. SEPIR system is still in a prototypical development and it has been used by only two experts. Hence, a qualitative, and statistically significant, evaluation to assess the quality of re-rank after the personalisation took place is not feasible at the moment. We decide to perform several queries with and without personalization, then by computing the distance between rankings, we measure how the personalization algorithm affects the order of documents in the ranking.

In order to simulate a large number of queries, we initially set up a query q_0 and then we automatically generate other queries taking into account the n words most related to q_0 computed in the *WordSpace*. Each related word represents a new query, then we have a total number of $n + 1$ queries. For each query we compute the distance between the ranking with and without the personalization. In order to simulate the user session (the user profile), at the time of query q_i we consider as a user profile all the previously issued queries: $q_0 \dots q_{i-1}$.

We perform the evaluation both in English and Italian by exploiting the collection of documents adopted in the use cases. In particular, we have two collections: the first (*Patent*) contains all the documents used in the patents analysis, while the second (*Project*) contains all the documents coming from the projects analysis.

For each collection, we select ten seed terms (q_0) exploiting the name of the technologies extracted in both the use cases. For each seed term, we consider the ten most related words for generating the other queries. For each collection, we have a total of 100 pairs of rankings to compare. We use the Kendall's Tau measure to compute the distance between rankings. Kendall's Tau values can range from -1 (opposite ranking) to 1 (same rank), with 0 meaning that there is no correlation between the two. The final value is the average of the distances computed for each pair.

We perform the evaluation by taking into account several top-N documents for each query, in particular: 10, 25, 50 and 100. Results are reported in Table 4.

Top-N	**Patent**	**Project**
10	0.708	0.731
25	0.736	0.695
50	0.744	0.652
100	0.738	0.616

Table 5 Kendall's Tau values between rankings with or without personalization.

From the table we can notice that in general the rankings are slightly different. It is important to underline here that the personalization method only re-ranks documents according to the user profile and it does not add/remove documents to the rank. Among

the two collection, the distance is more evident in the *Project* documents (lower values for 25, 50, and 100). This can be ascribed to the highest heterogeneity of documents in this collection. Indeed, the *Patent* collection has been built retrieving documents that are relevant for the same set of queries, hence with a high probability of similar content and a lower sensitivity to the re-ranking.

5 Related Work

In this section we firstly discuss the state-of-the-art on personalization techniques for information retrieval, then we show some related works on these methods for e-Government services.

5.1 Personalized Information Retrieval

One crucial point in personalised systems concerns how to gather, represent, and exploit information about the user in order to provide personalised services both on the Web and in enterprise environments [MGSG07, GSCM07]. A typical search process involves a user submitting a query, often as sequence of terms, to a search engine and receiving a ranked list of documents. A classic Information Retrieval (IR) system is based on the one-size-fits-all approach: the response of the system to the same query issued by different users is always the same. Conversely, a personalized IR models include the user model in the ranking formula: the concept of relevance is extended to the user interests [Sil10] so that different users can obtain different results for the same query.

The integration of personalization in a IR model consists of three phases: 1) collecting the information about the user preferences, 2) modelling this information in order to build a formal representation of the user profile and 3) integrating the user profile into the IR model. In the latter phase, different approaches can be used either to adapt the user query or to re-rank the results.

Among query adaptation approaches, Yin et al. [YSC09] proposed a query expansion technique that uses external evidence obtained from Web search engines to expand the original query. Queries and clicked documents are represented using a Query-URL graph on which a graph-based machine learning algorithm is applied. The Query-URL graph is a bipartite graph where the first set of vertices represents the queries, while the second represents the documents. The edges connecting the vertices of the two sets represent click-through information. A random walk algorithm applied on the graph generates the probabilities between queries: higher probabilities reflect higher query similarities. These similarities are then exploited to improve future searches by query expansion. Query log information is exploited by Cui et al. [CWNM03] where information about the user profile is aggregate as in a collaborative filtering approach. Other methods collect information about a individual user. Zhou et al. [ZLW12] use social media information to build a specific user model, while Chirita et al. [CFN07] use the data extracted from personal desktop documents, emails and cached Web pages to expand the initial query terms. Koutrika and Ioannidis [KI04] propose a rule-based query re-writing process for personalising structured search across a database of movies. The system replaces the submitted query with multiple queries using a set of rules based on the movie preferences of the user. In explicit relevance feedback approaches to query expansion, users are asked to explicitly provide feedback about the relevance of documents to their information need. This feedback can be either

positive or negative, for example by marking documents on a binary scale as relevant or non relevant. The system analyses the feedback and modifies the original query accordingly. The new query is then used to retrieve documents that are similar to the positive examples, or filter out documents that are similar to the negative examples. Ruthven et al. [RL03] propose a survey on the use of explicit relevance feedback methods.

Other common approaches to search personalization exploit the user profile for re-ranking or document scoring. Result re-ranking is the technique used in SEPIR system and it takes place after the retrieval of the set of relevant documents, when an additional re-ranking is performed to re-order documents on the basis of the user profile. Speretta et al. [SG05] proposed MiSearch, a system where documents and snippets retrieved by Google are passed to the re-ranking module. In this module the snippets are classified by their conceptual content with respect to Open Directory Project categories. After the concepts of the snippets have been deduced, they are compared to the concepts in the user model using cosine similarity. Then, the results are re-ranked in descending order of the conceptual similarity score. Rather than re-rank the initial retrieved set, result scoring approaches incorporate personalisation features directly in the ranking formula of the retrieval model. De Gemmis et al. [dGSLB08] learn a semantic user profile in a probabilistic model by using the explicit user feedback and the synsets of WordNet. Then, the user profile is introduced in the ranking formula with the user query.

5.2 Personalization and Semantics in e-Government services

The problem of information overload also involves the governance field. E-Government services, which aim to satisfy citizens, other institutions and entrepreneurs, should not be limited to the publication of information on the Web. The variety of actors (such as citizens, businesses, employees, local administrations and academia) that exploit e-Government services forms a complex network of users with different requirements and needs that interact with retrieval services in the quest for relevant information. In this context, personalized e-services and semantic methods are required in order to ease the user experience through the intelligent information access.

Gue and Lu [GL07] propose a recommender system called Smart Trade Exhibition Finder (STEF) for one-and-only items in e-government services in order to provide intelligent e-government services with personalized recommendation techniques. STEF combines semantic similarity on product taxonomies with a classical item-based collaborative filtering approach. Reiterer et al. [RFJ$^+$15] propose WeeVis, a constraint-based recommender that exploits the MediaWiki knowledge base. The authors suggest that WeeVis can be used in e-Government domain for the on-line advisory service for citizens, for modelling internal processes —like the signing of travel applications— or as an information platform with an integrated knowledge-based recommender for community residents.

In [BWK$^+$15] is described a method to provide personalized cultural heritage information in order to present personalized information to the user. This method collects the user information through a mobile museum guide. Semantic models and Linked Open Data are used to represent the regional assets as Cultural Objects. Then, user preferences are used to obtain relevant Cultural Objects, while some features are used to determine whether an event or a cultural heritage place is desired.

Biancalana et al. [BMS15] propose Personalized Extended Government (PEG), a retrieval model that simplifies and enhances the effectiveness of e-Government services in the context of G2G e G2C. Since the information in a local administration can be structured,

semi-structured or unstructured, a knowledge indexing component extracts and indexes the knowledge in a coherent way. In details, a NLP pipeline with a named entity recognition module processes the unstructured text. Then, a user profile represented by a concept network is used at query time to provide personalised results.

In [VdCFH14] the authors carry out a comparative evaluation based on the combination of three different user profile representations to support citizens access to the documents of the Andalusian Parliament. The first approach is a weighted concept profile, where the concepts represent abstract topics of interest for the user, the second one is a common user profile based only on the keywords, and the last is a hybrid approach that combines the first two ones. The authors show that the user profiles help the citizens find information relevant for their needs.

Social media and microblogging, like Twitter, have proved to be a valuable source of data to support the government decision processes such as emergency management [KPA09][Liu14][SOM10][KIS13], health-related matters [BTY12][LWS+10] and citizens engagement [WP14][PW11]. Since data from social media are provided as unstructured text, understanding citizens behaviour requires natural language processing in order to extract semantic information. TweetAlert, a tool that uses semantic components for citizen opinion mining, is proposed in [VCC14]. TweetAlert collects tweets and provides several semantic API for text classification, topic extraction and sentiment analysis to improve government services.

6 Conclusions

This paper described SEPIR, a SEmantic and Personalised Information Retrieval Tool for the Public Administration. SEPIR was developed for the public administration of Apulia Region in order to extract knowledge and strategical information from documents which support the decision making and the strategic planning in e-Government activities, in particular for e-Procurement purposes. We described the main functionalities of the proposed system, which provides personalized retrieval of relevant information and visualization tools. The personalised search engine has been implemented through the Random Indexing technique, which builds a semantic space where similar terms are represented close to each other. Then, the user model consists of a vector representation in this space of the terms exploited in the past user queries. Moreover, the weighting scheme of such terms reflects the recency of the query. This personalization technique was developed in two different retrieval models: the former is based on a classical vector space, while the latter makes use of Random Indexing to build a document space where similar documents are represented by close vectors. The visualization tools exploit such vector representation of documents to show a correlation matrix where similar documents are clustered together. Moreover, the key-phrases extracted from the collection are rendered in bubble charts as to give in a glance an overview of the most common phrases and concepts in the collection.

The potentiality of these techniques has been showed in two case studies involving both the Italian and English language. We prove the effectiveness of our system to support the decision making process in two scenarios: patent analysis for the e-procurement and funded projects analysis for measuring the impact of financed technologies.

As future work, we plan to conduct a more extensive evaluation involving a large number of regional experts as users. Moreover, some open challenges are still present. We need to improve the phrases extraction method in order to avoid no relevant concepts and we

are studying an approach for integrating knowledge extracted from textual content with structured knowledge already stored in the digital databases of the Apulia Region.

Acknowledgement

This work is supported by the project "Multilingual Entity Liking" funded by the Apulia Region under the programme FutureInResearch and by the project "A NLP framework for the analysis of Italian documents" funded by InnovaPuglia S.p.A.

References

[BMS15] Claudio Biancalana, Alessandro Micarelli, and Giuseppe Sansonetti. Personalized extended government for local public administrations. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 23rd Conference on User Modeling, Adaptation, and Personalization (UMAP 2015)*, volume 1388 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[BTY12] Jiang Bian, Umit Topaloglu, and Fan Yu. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, SHB '12, pages 25–32, New York, NY, USA, 2012. ACM.

[BWK+15] Antonino Lo Bue, Alan J. Wecker, Tsvi Kuflik, Alberto Machì, and Oliviero Stock. Providing personalized cultural heritage information for the smart region - A proposed methodology. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 23rd Conference on User Modeling, Adaptation, and Personalization (UMAP 2015)*, volume 1388 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[CFN07] Paul Alexandru Chirita, Claudiu S. Firan, and Wolfgang Nejdl. Personalized query expansion for the web. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 7–14, New York, NY, USA, 2007. ACM.

[CWNM03] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. Query expansion by mining user logs. *IEEE Trans. on Knowl. and Data Eng.*, 15(4):829–839, July 2003.

[DG99] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. Technical report, TR-99-006, International Computer Science Institute, Berkeley, California, USA, 1999.

[dGSLB08] Marco de Gemmis, Giovanni Semeraro, Pasquale Lops, and Pierpaolo Basile. A retrieval model for personalized searching relying on content-based user profiles. In *6th AAAI Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP 2008)*, pages 1–9. Chicago, USA, 2008.

[DL05] Yi Ding and Xue Li. Time weight collaborative filtering. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 485–492, New York, NY, USA, 2005. ACM.

[Fan02] Zhiyuan Fang. E-government in digital era: Concept, practice, and development. *International Journal of the Computer, The Internet and Management*, 10(2):1–22, 2002.

[GL07] Xuetao Guo and Jie Lu. Intelligent e-government services with personalized recommendation techniques: Research articles. *Int. J. Intell. Syst.*, 22(5):401–417, May 2007.

[GSCM07] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. User profiles for personalized information access. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web, Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, pages 54–89. Springer, Berlin, Heidelberg, 2007.

[Har68] Zellig S. Harris. *Mathematical Structures of Language*. Interscience Publishers, New York, 1968.

[Kan88] Pentti Kanerva. *Sparse Distributed Memory*. MIT Press, Cambridge, MA, USA, 1988.

[KI04] Georgia Koutrika and Yannis Ioannidis. Rule-based query personalization in digital libraries. *Int. J. Digit. Libr.*, 4(1):60–63, August 2004.

[KIS13] Karl Kreiner, Aapo Immonen, and Hanna Suominen. Crisis management knowledge from social media. In *Proceedings of the 18th Australasian Document Computing Symposium*, ADCS '13, pages 105–108, New York, NY, USA, 2013. ACM.

[KPA09] Kirill Kireyev, Leysia Palen, and Kenneth M. Anderson. Applications of Topics Models to Analysis of Disaster-Related Twitter Data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada, 2009.

[Liu14] Sophia B. Liu. Crisis crowdsourcing framework: Designing strategic configurations of crowdsourcing for the emergency management domain. *Comput. Supported Coop. Work*, 23(4-6):389–443, December 2014.

[LWS+10] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, pages 117–125, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[MGSG07] Alessandro Micarelli, Fabio Gasparetti, Filippo Sciarrone, and Susan Gauch. Personalized search on the world wide web. In *The Adaptive Web, Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, pages 195–230, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., USA, 2013.

[PW11] Cecile Paris and Stephen Wan. Listening to the community: Social media monitoring tasks for improving government services. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 2095–2100, New York, NY, USA, 2011. ACM.

[RFJ+15] Stefan Reiterer, Alexander Felfernig, Michael Jeran, Martin Stettinger, Manfred Wundara, and Wolfgang Eixelsberger. A wiki-based environment for constraint-based recommender systems applied in the e-government domain. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 23rd Conference on User Modeling, Adaptation, and Personalization (UMAP 2015).*, volume 1388 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[RL03] Ian Ruthven and Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, June 2003.

[Sah05] Magnus Sahlgren. An Introduction to Random Indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, volume 5, 2005.

[SG05] Micro Speretta and Susan Gauch. Personalized search based on user search histories. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, pages 622–628, Washington, DC, USA, 2005. IEEE Computer Society.

[Sil10] Fabrizio Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1–2):1–174, 2010.

[SOM10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[SWY75] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[VCC14] Julio Villena-Román, Adrián Luna Cobos, and José Carlos González Cristóbal. Tweetalert: Semantic analytics in social networks for citizen opinion mining in the city of the future. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014.*, volume 1181 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[VdCFH14] Eduardo Vicente-López, Luis M. de Campos, Juan M. Fernández-Luna, and Juan F. Huete. Personalization of parliamentary document retrieval using different user profiles. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014.*, volume 1181 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[WP14] Stephen Wan and Cécile Paris. Improving government services with social media feedback. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, pages 27–36, New York, NY, USA, 2014. ACM.

[YSC09] Zhijun Yin, Milad Shokouhi, and Nick Craswell. Query expansion using external evidence. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 362–374, Berlin, Heidelberg, 2009. Springer-Verlag.

[ZLW12] Dong Zhou, Séamus Lawless, and Vincent Wade. Improving search via personalized query expansion using social media. *Inf. Retr.*, 15(3-4):218–242, June 2012.