Aid to Regional Development Agencies: Finding and Matching Research Funding Opportunities

Michael Kaschesky
Bern University of Applied Sciences
E-Government Unit
Bern, Switzerland
+41 31 848 3433
ksm1@bfh.ch

Guillaume Bouchard
Xerox Research Center Europe
Data Mining and Machine Learning
Grenoble, France
+33 4 7661 5198
guillaume.bouchard@xrce.xerox.com

Stephane Gamard
salsaDev SA
Chemin de Aulx 18
Plan-les-Ouates, Switzerland
+41 22 884 8650
stephane.gamard@salsadev.com

Adrian Gschwend
Bern University of Applied Sciences
E-Government Unit
Bern, Switzerland
+41 31 848 3440
gea1@bfh.ch

Patrick Furrer
euresearCH
Effingerstrasse 19
Bern, Switzerland
+41 31 380 60 04
patrick.furrer@euresearch.ch

Reinhard Riedl
Bern University of Applied Sciences
E-Government Unit
Bern, Switzerland
+41 31 848 3440
rer2@bfh.ch

ABSTRACT

Regional development agencies are confronted with a plethora of research funding programs that provide opportunities for regional research groups and SMEs. This paper describes the practical benefits and technological building blocks of an online matching service of research profiles with funding opportunities. It gives an overview of the infrastructure for data sourcing and processing based on a scalable cloud computing platform. It then describes how data is consolidated, analyzed, and interlinked so as to facilitate the finding and matching of funding opportunities. Integrating concepts and themes with those available as Linked Open Data (LOD) adds value to finding and matching by interlinking results with publicly available data resources, and to improve search performance for related content on the internet. In order to optimize matching, the paper describes the user-profiling capabilities based on statistical analyses and machine learning. The paper concludes by discussing the lessons learned so far as well as possible extensions into other application areas.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing
H.3.1 [**Content Analysis and Indexing**]: Linguistic processing
H.3.3 [**Information Search and Retrieval**]: Information filtering

General Terms

Management, Experimentation, Human Factors, Verification.

Keywords

Measurement, Design, Economics, Reliability, Standardization.

dg.o '12, June 04 - 07 2012, College Park, MD, USA
Copyright 2012 ACM 978-1-4503-1403-9/12/06 $10.00.

1. INTRODUCTION

European research and innovation programs are becoming more complex to understand and to access due to the high fragmentation of the European research area and ever new types of funding types and programs with sub-programs. The road to Framework Program 8 (FP8) is not going to bring much simplification, with the European Innovation Partnerships, the Joint Programming Initiatives, the Knowledge and Innovation Communities, the Joint Technology Initiatives, the coordinated calls (bilaterally with international partner countries), the increasing number of Integrated Projects and ERA-Nets which themselves launch calls for proposals.

In addition to this European fragmentation, each member state and associated state develops their national and even regional research and innovation programs. This makes the public funding space increasingly harder for regional development agencies and researchers to understand and access. Even for national contact points, whose mission is to keep researchers informed about these funding opportunities, the task is becoming ever more complex.

In addition, more and more countries around the world tend to open their research, innovation and education programs to international participation, in particular the emerging economies. The breadth of new opportunities for researchers, students, and innovators is literally proliferating and accelerating every day, and the time to access and learn about them is not extensible enough, especially for SME clients, to ensure the proper transparency about these opportunities.

Researchers and SMEs willing to participate in public funding programs, and FP7 specifically, have to find the right sub-programs, the right call, the right topic, and the right partners. This is a very challenging task as the programs can be very complex, including in addition to the thematic opportunities different horizontal programs being of potential interest to SME like the Industry Academia Partnerships and Pathways (IAPP) fellowships in the FP7-PEOPLE sub-program or the SME specific measures in the FP7-Capacities sub-program. The Enterprise Europe Network links to existing competencies but, again, SMEs have to scan the database manually or to subscribe by using a keyword based profile.

Euresearch, the Swiss national contact point for research and innovation programs, has collaborated with salsaDev and Bern University of Applied Sciences in order to develop an innovative recommender system for consolidating, managing, matching and distributing information about public funding opportunities (e.g. FP7 and related EU programs, as well as national programs) and innovation partnership collaborations (e.g. Enterprise Europe Network). The research concerning the expansion of this initial concept with machine learning algorithms to improve matching of research groups and SMEs to funding opportunities is the focus of this paper.

The automated and user-adaptive matching of research groups and SMEs to funding opportunities is the key advantage of the ICT-based knowledge system vis-à-vis manual or existing ICT-based approaches. Therefore, a key asset is information provided by the user, either indirectly through user behavior or directly through user contributions such as crowdsourcing or profile uploading. Data collected from the user constitutes the basis for intelligent finding and matching of funding opportunities to user-specific needs concerning research and development. Information provided by the user enables ever more precise matching results leading to increased user engagement thereby setting off positive network effects towards increasing user adoption. Hence, the core value proposition is summarized as follows: More relevant (correct and specialized) matching results because accuracy (correctness) improves with quantity and quality of user data while variety (breadth) increases with number of users.

The technical innovations that make intelligent finding and matching possible focus around the user-adaptive system [1]. A user-adaptive system refers to the ability of an ICT system to monitor and learn the specific needs and preferences of a user and align its invoked features, functionalities, and graphical interfaces accordingly. It enables more usable and customized interfaces and improved performance due to user-aligned prioritization of tasks based on machine learning and collaborative rather than rule-based approaches. The user-adaptive system aligns to user needs and preferences and provides suggestions based on user activity and implicit feedback signals as well as on explicit feedback and collaborative techniques (such as crowdsourcing).

This paper is structured into six sections including this introductory section 1. In section 2, we describe the prototype and examples in order to frame this paper and introduce the technological challenges that this paper addresses.

Section 3 provides an overview of the infrastructure for data sourcing and processing, which is based on a scalable cloud computing platform. Offering applications as services instead of installing the software on-site has the advantage that any interested user can sign-up and use the software without installing it. Because load (e.g. number of concurrent users, number of data sources) can vary and may spike at unforeseen times, a scalable cloud computing platform is a requirement for data processing services.

In section 4, we describe how data is consolidated, analyzed, and interlinked so as to facilitate the finding and matching of funding opportunities. For example, data from the Cordis FP7 database with information on the European research funding details is available as Linked Open Data (LOD) as are many other public databases. Integrating retrieved concepts and themes with those available as LOD adds value to finding and matching by interlinking results with related LOD resources, and to improve search performance for related content on the internet.

In section 5, we describe the user-profiling capabilities aimed at optimizing matching results based on statistical analyses. For example, this enables the “suggested opportunities” feature and the ability to confirm and disconfirm results, which are important to optimize the accuracy of results for the user. Optimization is semantically calculated based on the elements available for reasoning, such as metadata (e.g. type of user, country), implicit user inputs (e.g. previously browsed opportunities), and explicit user input (e.g. confirmed opportunities). In a Bayesian framework, these optimizations are made possible thanks to statistical learning techniques through three distinct modeling steps, an a priori probability (i.e. the initial matching results), new observational data (i.e. metadata and user input), and the a posteriori distribution which is used to optimize matching in the next round (i.e. optimized matching results).

Finally, in section 6, we provide an overview of the lessons learned so far as well as possible extensions into other application areas, such as technology intelligence for regional development agencies or commercial firms.

2. METHODOLOGY AND EXAMPLES

The underlying methods for matching content are based on text similarity matching algorithms developed by salsaDev. The salsaDev matching service is context sensitive. That is, each document has a set of parameters attached to it and it is possible to apply a selection of the documents satisfying the constraint before running the actual matching algorithms. For example, one parameter could be the EEN type (technology offer, request, business offer, request or partner search). A constraint could limit the search of opportunities to the technology offers. The search would be then restricted to these opportunities. Constraints are expressed as key-value pairs and provide the functionality to execute the search within a sub-set of the index.

it is not within the scope of this paper to describe the underlying methods, rather, to describe optimization methods, especially using user profiling, that provide the functionality needed for the end user, in this case a regional development agency.

Euresearch is currently facing the following difficulties:

- Euresearch supports its clients not only for FP7 or EEN but also for many other programs (COST, ERA-Net, IMIs, etc.). Each program comes with its own specific classification structure (Health, ICT etc.) and makes it impossible to manually profile the client in all these schemes.
- The past experience shows that it is difficult to motivate the client to profile her/himself over keywords. The opportunity finder provides a simple way to gather clients interest by investigating the selected opportunities. The use of supplementary information (selected opportunities, free text) for the profile and its benefit has to be investigated.
- The client receives currently e-mails based on his/her FP7 profile (keyword based). The existing FP7 keyword list has to be assessed and possibly modified and improved.

2.1 Optimized content matching using user profiling techniques

The user profiling makes extensive use of search and categorization. This section gives a short description of the basic ideas behind the profiling.

In information retrieval the task is, given a query q and a set of documents (or opportunities in our case) {d1,d2,...,dN}, to rank the documents according to their similarity to q, the highest rank providing the most relevant documents. If a user profile can be treated as a query, promoting content based on a user-profile is equivalent to find the documents which best matches the profile.

Unfortunately, most of the client's profiles are not expressed in terms of a query q, but in terms of categories. Typically a person’s interests are expressed as a set of categories {e.g. “Safety”, “Environment”, “Waste Management”}. The categorization/ classification problem aims at assigning a record (document or opportunity or person's profile etc.) to one or more categories within a predefined set of categories {c1,c2,...,cM}.

In our case the profile is the combination of up to three elements:

- Categories: a set of categories of interest to the user. These categories are expressed in a given taxonomy.
- Opportunities: a collection of opportunities of interest to the user (similar to bookmarking of opportunities).
- Free text: a raw text query (keyword or sense) describing the user's interest.

These three elements will be combined as illustrated in the table below. Each diagram represents a pseudo “document-index” space. The document index or categorization space is generally a high dimensional space (50 to 1500 dimensions) but is illustrated here in a 2-dimensional space with terms 1 and 2.

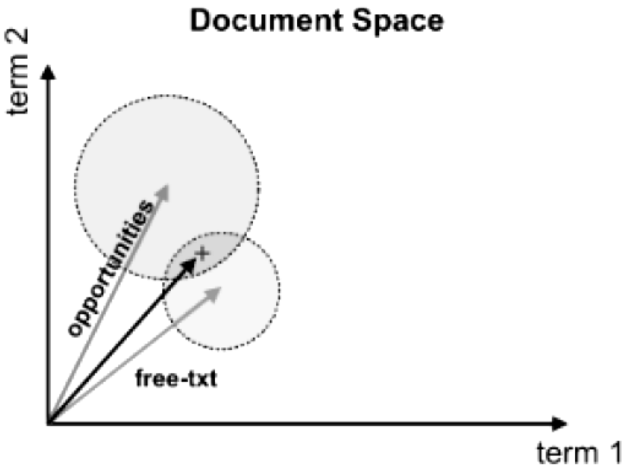

Figure 1: Components in the document-index space with lower vector computed from user profile and upper vector based on opportunities where overlap defines the region containing opportunities of interest

SalsaDev provides a search service capable of querying its index based on either raw text (used for user-profile query) or a set of opportunities (pre-indexed content - currently used in the “similar” content widgets of the Opportunity-Finder). Figure 1 above shows both profile components in the document-index space: the lower vector is computed from the free-text of the profile, the upper vector is given by the list of bookmarked opportunities and the overlap of both defines the region containing opportunities of interest for the respective components of the profile. While it is impossible to merge both types of queries both result sets are comparable and space equivalent, i.e. they can be merged at no additional computational costs.

2.2 Semantic Similarity Matching

There are 3 major steps to provide results based on a user profile.

- Information Access (search): By executing a search based on the user's bookmark and its query we are creating a first sub-set of the index which is most relevant to the user's interest. It is important to keep in mind that this subset can be the whole index itself (although the result set will be ranked accordingly to query and bookmarks). This information access step (which is 2 merged searches) can also be used to constrain the set from which the recommendations will be made (constrains).
- Aggregation (merge): this step is more an incremental set, preparing for the final step. During merger heuristics drive the weighting of independent searches. Weights can be applied for constrains (are news more important than blogs for example), for metadata (i.e. latest news rather than old ones) and sense-heuristic (is the query a set of keywords or a complete CV?)
- Evaluation: The final set, and most computationally expensive, of the recommendation process. The evaluation step transforms the result set generated by the information access process by ranking its aggregated results with respect to the categories of the profile. There are a couple of key points such as a) evaluation and information access are completely independent with respect to the taxonomy; b) not the raw score of categorization ranks the result, but the computed intersection (relevancy in each process) of the fit; c) even if information access fails, evaluation will sort the constrained set accordingly to categorization scores.

The first example concerns the semantic search which utilizes the text similarity matching algorithms in order to compare and retrieve text with the similar meaning. In this case, the user inputs a paragraph or more text, for example:

« While flexible pressure and temperature sensors are relatively easy to realize, gas sensors are more challenging due to their more complex structure involving gas sensitive materials that must have access to the external environment. Particularly for gas sensors targeting smart textile applications, the challenge is significantly greater since the industrial weaving is a very abrasive process. The main target of TWIGS is to design, fabricate and characterize a mechanically robust thin-film capacitive gas sensor platform compatible with industrial weaving processes. The targeted sensor platform is particularly interesting to develop textile integrated sensors arrays with each individual sensor functionalized with a different sensing material to perform complex sensing tasks »

The list of results is retrieved as described above. By selecting the first opportunity (e.g. clicking on the title) one reaches the detailed description of the opportunity (in that case a project description of an FP7 project under the Food, Agriculture and Biotechnology theme). On side panels, additional opportunities that are closely related (semantically speaking) with that opportunity are presented, to allow further exploration. Scientific officers at Euresearch confirm that they would have advised a client to look into this same opportunity, given the original query the client had formulated.

It is also possible to use a few keywords as in most internet search engines to specify the query, such as “smart textiles“.

In this case, the Opportunity Finder proposes a technology opportunity in «conductive ink for smart textiles», a business opportunity in «trade intermediary services for smart casual wear» and one technology request in «cosmetotextile lotion».

After selecting via filtering only those R&D sub-programs in nanotechnology, the list of opportunities is completely changed and focuses now specifically on opportunities related to the NMP program (where it really belongs to), and further down in the list to the ICT program (smart components and integrated systems).

The remaining sections describe the procedural steps, not the technical details, to accomplish the opportunity matching. The procedural steps serve to illustrate and explain the overall approach. It is, however, not within the scope of this paper to detail the technical specifications, methods, and algorithms underlying each of the steps.

3. DATA SOURCING AND PROCESSING

The first part of this section describes data bootstrapping and data handling. The system accesses and indexes internal and external content through web harvesting and structured data sources. The process of data curation is fully automated for structured data sources, such as patent databases and Linked Open Data (LOD), and semi-automated for unstructured data sources, such as online reviews or comments. The process results in high-level abstraction and persistence of the data for reuse based on shared metadata models. In this way, the automated collection, consolidation, quality-assurance, and persistent availability of data ensures greatly minimized manual intervention when gathering and indexing content.

The second part of this section describes the strategies and methods for preserving privacy when gathering and analyzing data. Data on user interactions remain completely anonymous and even when users upload content only an abstracted index is stored for enabling instant interlinking at the desired level of publicity (e.g. team, firm, partners, public). The privacy-by-design framework will allow associating public and private data without compromising the privacy of the users.

The third part of this section describes the scalable and high-performing computing infrastructure designed for harvesting, mining, and matching large amounts of data. The cloud computing platform to be implemented enables on-demand access by many different organizations on a per-usage billing (or free for basic services). Cloud computing enables rapid scalability because resources, such as computing capacity and storage, or services, such as databases or messaging, can be rapidly accessed and exited based on the current demands of users. Rapid scalability is necessary to accommodate the computationally and data intensive machine learning methods of the user-adaptive system.

3.1 Data gathering and indexing for high-level data abstraction and persistence

Two types of data sources are pooled in the system: unstructured and structured data. Unstructured data refer to data that must be processed before it can be correlated and matched with existing data. It may be obtained from natural language content, for example, through text feature extraction. Data sources exhibiting unstructured data may include, among others, web pages on the internet or text chunks stored in patent databases. In contrast, structured data refer to data that can be directly correlated and matched with existing data after being retrieved. Structured data can be divided into those available through traditional databases (via SQL) and those available through the many LOD sources (via SPARQL).

Within this project, access to many structured data sets relevant for the purpose is available. For example, the Fairview Alexandria patent database contains roughly 85 million patent records from more than 70 countries with one of the largest corpus of full text records. GoodRelations is an example of LOD sources accessible via SPARQL endpoints. It provides a standardized vocabulary for product, price, and company data to be embedded into and processed by the applications developed in this project. In addition, the entire LOD data space can be accessed for pooling, such as publication databases (e.g. ACM, CiteSeer, DBLP, IBM, IEEE).

In order to provide comprehensive results for the end-user, correlation and matching is performed over a wide variety of unstructured and structured data from predefined sources (e.g. FP7 Cordis, Fairview Alexandria, GoodRelations, CiteSeer) and from emerging sources. Emerging sources are those that have been added from internet search using the Google Custom Search API. Based on user requirements, a large number of potentially relevant websites (e.g. cordis.europa.eu/fp7 or ted.europa.eu) are predefined in the Google Custom Search and are emphasized when searching the entire web. In addition, relevant other content is also included from sources which may not be known beforehand (e.g. publications, news, discussions).

Underlying data gathering is a robust data model that combines fine-grained metadata, quality controls, and data indexing into a well-defined data structure. Basic data column (attribute) metadata are generated automatically when data are loaded, but comprehensive metadata can also be pre-defined and saved as reusable metadata templates. Templates contain boilerplate documentation, detailed data column descriptors (e.g. name, units, description, data type, precision), and Q/C rules for each column. When templates are applied, data are validated and Q/C rules are applied automatically. Multiple templates can be defined for each data source to support different reporting standards or vocabularies.

Whenever possible, metadata and descriptors are stored rather than the original content. The link provides access to the originating repository. Original content is indexed and abstracted using a schema-free key-value (enabling actionable facets). Thus, storage mainly involves storing metadata and abstracted summary representations enabling lightweight access and faster analysis.

Persistence for CMS-like usage (e.g. archiving, browsing) and for text mining, matching, and Latent Semantic Analysis is achieved by two-step process. The first step is applied on natural language content and yields so-called “first-level semantics”, that is, vectors and ontology-based categories that describe the content. The second step involves the generation of semantic triplets describing the subject-predicate-object-relationship of entities within the text, so-called “second-level semantics”. Triples are most appropriate to store the extracted information independent of a specific use case or application and are widely deployed in

Semantic Web approaches using the Resource Description Framework (RDF).

3.2 Privacy preserving data gathering and analysis

The core benefits concern the matching and interlinking of data with Semantic Web resources and transfer learning from individual user behavior. These benefits can only be realized in a privacy-compliant way that accounts for the various privacy regulations in different jurisdictions. The technological challenge concerns the development of anonymization methods that will allow data fusion from public and private sources without endangering the privacy of the users. Anonymization methods preserve the privacy of the users so that no third party will be able to infer additional information.

A naïve anonymization method consists of the removal of direct identifiers like the name of a person or its social security number. These methods remain vulnerable to inferences based on indirect data that may identify the real person. The anonymization methods to be developed will protect against direct and indirect inferences by transforming indirect data. This transformation must balance two requirements:

- Maximize privacy by accounting for complex combinations of potentially identifying data;
- Minimize transformations of indirect data to maintain system accuracy and responsiveness.

The assignment of user interaction data with users is safely encrypted and stored and is used only for recommendations to the user itself and constraint to the purposes of the system (e.g. not for third party marketing). User interaction data that is not assigned to a specific user is also not be visible to other users or third parties, but it is aggregated with other user interaction data for making user-specific recommendations.

3.3 Cloud computing platform with minimal implementation costs

Software-as-a-service (SaaS) provides singular software applications (e.g. CRM) which are independent from each other. In contrast, infrastructure-as-a-service (IaaS) provides just infrastructure instances where the customer is responsible to acquire and maintain software applications and the logic between them. Particularly concerning the logic between applications, both SaaS and IaaS lack critical features that would enable cloud computing to achieve the interoperability and performance of current enterprise computing. Platform-as-a-service (PaaS) is the missing link that serves as native application infrastructure for cloud-based applications (SaaS) and controls the sharing of hardware resources (IaaS) between logically isolated applications (multi-tenancy), the scaling of applications to many users or use instances (elastic scalability), and the monitoring and management of application resources (scheduling).

PaaS can be taken further using Complex Event Processing (CEP) as described recently [4]. Compared to existing approaches (e.g. batch processing in Hadoop), CEP becomes relevant for analyzing large amounts of incoming data and deriving results from it to be delivered to a user or another application. Each input data constitutes an event within an event stream and within the ‘cloud’ of events in which various filtering, correlating, aggregating, and computing methods are applied to detect complex events (interrelated, “connect the dots”). The definition of an event producer describes the events it produces while the definition of an event consumer is defined in terms of the events it consumes. In a cloud-based approach, CEP can be used to distribute analysis across multiple computing nodes located near the various data sources resulting in performance gains. CEP improves PaaS because it now includes the processing and analysis of events (event-driven). For example, application developers can use various complex event types in their software without worrying about the underlying CEP engines.

In order to enable web harvesting, organization mining, and the advanced matching services, an event-driven cloud platform is designed and realized based on existing cloud computing offerings. For example, the Amazon Simple Notification Service (Amazon SNS) facilitates the creation of event-driven processing and messaging based on existing Amazon cloud EC2 (IaaS) and Elastic Beanstalk (PaaS). Apache Hadoop can run Amazon cloud offerings and is particularly suited for web harvesting, text processing, machine learning, and advanced data analytics. It uses batch processing to do data-intensive work in parallel. Red Hat’s new OpenShift PaaS is an open-source option supporting different programming languages and frameworks. These existing offerings greatly reduce development time and costs and provides an enterprise-ready cloud environment for the reliable and sustainable provision of the new services.

4. INFORMATION MINING AND INTERLINKING

The first part of this section describes content matching methods applied to indexed data, particularly through text mining. Text mining concerns deriving and extracting more aggregate features from text sources through the structuring of input data using parsing and extraction of linguistic features and the deep analysis and pattern identification of text data. Typical activities include text categorization, text clustering, entity (concept) extraction, and the production of granular taxonomies. The goal is to disambiguate information and to obtain more abstract high-level information in order to facilitate reuse of content.

The second part of this section describes processes of interlinking related content based on extracted text features and re-formatting as Linked Open Data (LOD). LOD concerns publishing structured content in reusable format so that text data can be directly interlinked and reused by others without parsing and text feature extraction. It makes use of identifiers (URIs), standard formats (RDF/XML), and included links and relationships to related texts and other objects. The goal is to closely combine text mining with LOD in order to provide advanced services over linked content so that users are able to utilize the best of both approaches.

4.1 Text mining and extraction for disambiguation and high-level abstraction

Text mining typically starts with retrieving a collection of text documents. This can be a dataset, single source, or composition of data from multiple sources. In order to make text documents available for computation, they encoded into numerical signatures. The standard representation in text processing for text analytics is known as vector space model [9]. It is usually implemented using a pipeline structure involving four steps: Tokenization (text is broken up into component words), Stop word removal (e.g. ‘the’), stemming (reducing words to their root form), and vectorization (numerical representation for texts) [6].

Additional steps include vector normalization, which enables the comparison between longer and shorter documents. For some applications, term weighting is performed. Term weighting has the effect of making words that are specific to a subset of documents more important. The natural language processing pipeline, which prepares the vector-space representation of documents for clustering and classification, uses the following additional processing steps: named entity recognition; co-reference resolution and entity normalization; and named entity disambiguation [6].

Named entities (e.g. names of people, organizations, locations, addresses) can be regarded as special tokens which often span several domains but act as single semantic units (e.g. "President Barack Obama", "IBM"). When enlisting the token types of a document (to translate it into coordinates in vector space), it is important to regard named entities as units. The named entity recognizer (NER) identifies the boundaries (starting and ending tokens) of named entities, and assigns them to categories. We have developed sophisticated NER technology for recognizing about 21 different categories in English and other major languages.

In addition, co-reference resolution (CR) algorithms identify groups of names in a document that refers to the same real world entity (e.g. "President Obama", "Barack Obama", "he"). This can be applied to identify all different mentions of an entity. The different forms can all be replaced (normalized) to a user-specified canonical form, which greatly facilitates the identification and retrieval of named entities (e.g. the instances of "Unites States", "U.S.", "USA", are all variants or synonyms of the same entity). We have developed tools for performing co-reference resolution for the most common entity types and solutions for efficient treatment of name catalogues that enable entity normalization.

Finally, named entity disambiguation (NED) helps in cases where a name can have several references, often belonging to the same entity type (e.g. "George Bush" may either refer to George Bush senior of junior). NED algorithms use the available context to hypothesize the most likely interpretation of ambiguous names.

4.2 Crowdsourcing and supervised automation in advanced data curation

The machine learning methods use supervised classification and active learning approaches for classifying text items, not just entire documents. In addition, methods for crowdsourcing, i.e. involving many end users, enable the system to recognize the specific taxonomies employed by users and to learn from user contributions. The multi-task learning algorithms enable the system to classify new text items automatically based on existing classifications of similar texts. These machine learning and crowdsourcing methods for taxonomy development and maintenance minimize manual intervention in data curation.

The relevant concepts, relations, and instances are defined in close cooperation with the end users. Concepts constitute the generic nodes in a semantic network, for example, person and place on a very abstract level. Relations define the various ways the concepts are interlinked, for example, a person *lives* in a place. Finally, concepts and relations are filled with real-world instances. There is a wealth of existing definitions and taxonomies that can be integrated for these purposes, in product development for example, the eClassOWL Web Ontology for Products and Services. The goal is to have light-weight ontologies describing the concepts, relations, and instances for each domain or scenario in order to expedite automated correlation and matching.

Ontologies and their taxonomies contain several thousands of entries, so automated tools to support the creation and maintenance of annotations and efficiently annotating content are needed. Supervised classification naturally fits this task. The user annotates content with own annotations and once the classifier learned how content and annotations are matched, the system suggests to the user similar content that is automatically annotated with similar annotations. If the user confirms, the system continues to annotate the type of content with that set of annotations. The more confirmations the system receives, the higher the accuracy of the resulting annotations [7]. These tasks have been extensively researched and standard baseline methods based on Multiclass Support Vector Machines or multinomial logistic regression (a.k.a. Maximum entropy classifiers) can be used.

4.3 Reusable content formats for automated interlinking, finding and matching

Extracted content and semantics need to be put in context to increase resulting value and provide unambiguous results. While hyperlinks are known to be valuable source of information for search engines [2], they might be missing between documents, e.g. a text might mention the name of an organization which is well described in another document. The resultant terms and knowledge need to be interlinked with same and similar terms/knowledge in the internet:

- Enrich the extracted content with existing information available in the Internet/Semantic Web;
- Interlink as much information as possible to increase the value of knowledge extraction;
- Use available public sector resources in Semantic Web and LOD format.

To combine text mining and content matching with public data sources, we use semantic web technologies. Many current and past ICT FP7 projects use RDF as its data model and started to bootstrap the web of data. In 2007, the LOD Cloud diagram contained just a handful data sets with a still countable amount of triples. In 2011 it is impossible to count all the available triples mirroring the exponential growth in the early days of the web.

The power of the Semantic Web lies in two simple things: A simple, scalable data model combined with links (or URIs, Uniform Resource Identifiers) to uniquely describe and identify knowledge. The data model is using a simple subject-predicate-object combination to form so-called triples. RDF is the propagated data model for triples in the semantic web and makes it possible to store this knowledge in both human and machine usable form. The fact that machines can access this knowledge makes it possible to perform tasks on huge amounts of data which otherwise would have to be carried out by human experts.

The real power of LOD lies in interlinking data sets with each other. DBPedia for example exposes all facts that can be exported from a Wikipedia entry in RDF, while ProductDB gathers data about any kind of product. ProductDB links to DBPedia whenever possible, as a result, information about the same thing can be derived from two different data sources. For that, a unique URI describing a thing exists in both data sources, which in itself serves as the interlink between the data sources.

5. KNOWLEDGE FINDING AND MATCHING

The first part of this section describes methods for pooling and finding content, which integrate the previous processing steps into a stream-lined application for searching and finding. This builds on the text-feature and entity extraction as well as the automated interlinking with LOD-formatted sources on the web and enables the refinement and filtering of search results based on integration with such metadata.

The second part of this section describes tools for correlating and matching content through the semantic matching engine. The matching engine extracts valuable contextual relationships between all content indexed by the system. This enables the user to view instantly a list of suggested content based on the user and profile or on the content that is currently viewed. The list of suggestions is grouped by categories to allow browsing or refining the results by category. Specific user-adaptive methods for collecting user feedback and activity enable the system to automatically learn and improve results over time.

5.1 Methods for finding, evaluating, matching and integrating content

The system features a powerful search engine enabling users to gain quick access to relevant data throughout the system. A key component of the search engine is its ability to refine and filter search results based on its integration of metadata generated in the previous steps. The search engine is powered by Lucene/Solr, the leading open-source search engine developed by the Apache foundation. In addition, other semantic engines are plugged to integrate results through RESTful API services. To make Lucene/Solr scalable and flexible based on the cloud platform described above, additional components are integrated (e.g. Infini, ehCache, JGroups). The key search-oriented features include:

- Search through all content harvested in the data pool;
- Faceted search (based on content categories, metadata and entities);
- Integration of Linked Open Data (LOD) to extend and refine search results;
- Cross-lingual indexing and search results (cross-referencing between languages);
- “Did you mean?”-Functionality in case of typos or spelling mistakes;
- Auto-completion of search queries.

The user’s search experience is critical for the success of learning methods and overall acceptance. The search engine indexes and integrates all data within the data pool while being specialized to the user’s needs (e.g. user profiling) so that relevant search results are quickly retrieved. The engine’s specialization is achieved through the integration of LOD, and the integration of categories, other relevant metadata, and extracted entities matched to the user. The specialization is a crucial component of the differentiation between the fully horizontal, traditional full-text search engines and enables query intent discovery ([3], part 1.2).

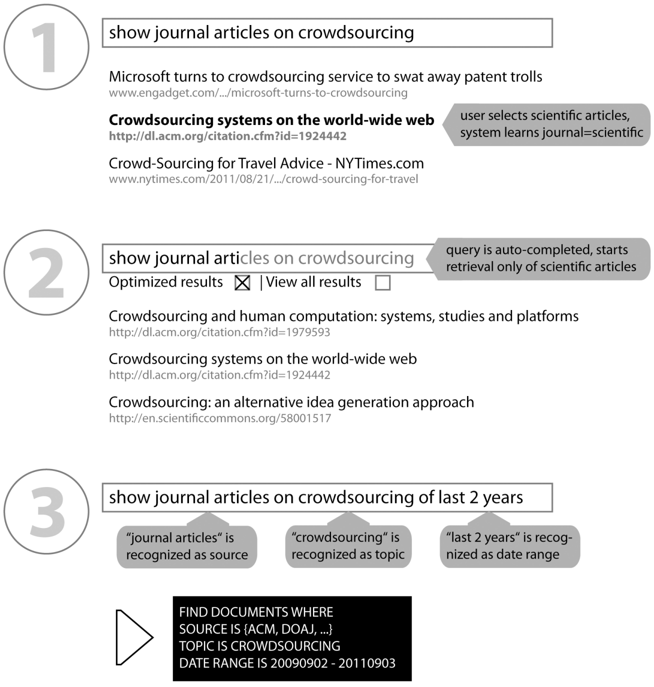
1
show journal articles on crowdsourcing
Microsoft turns to crowdsourcing service to swat away patent trolls
www.engadget.com/.../microsoft-turns-to-crowdsourcing
Crowdsourcing systems on the world-wide web
http://dl.acm.org/citation.cfm?id=1924442
user selects scientific articles, system learns journal=scientific
Crowd-Sourcing for Travel Advice - NYTimes.com
www.nytimes.com/2011/08/21/.../crowd-sourcing-for-travel
2
show journal articles on crowdsourcing
query is auto-completed, starts retrieval only of scientific articles
Optimized results ☒ | View all results □
Crowdsourcing and human computation: systems, studies and platforms
http://dl.acm.org/citation.cfm?id=1979593
Crowdsourcing systems on the world-wide web
http://dl.acm.org/citation.cfm?id=1924442
Crowdsourcing: an alternative idea generation approach
http://en.scientificcommons.org/58001517
3
show journal articles on crowdsourcing of last 2 years
“journal articles“ is recognized as source
“crowdsourcing“ is recognized as topic
“last 2 years“ is recognized as date range
FIND DOCUMENTS WHERE
SOURCE IS {ACM, DOAJ, ...}
TOPIC IS CROWDSOURCING
DATE RANGE IS 20090902 - 20110903

Figure 2: Simple search (1), auto-completion of search queries (2), and query intent discovery (3)

5.2 Methods for correlating and matching content

The intelligent search described above provides a quick and efficient way for users to search, refine and filter through relevant data pooled by the system. In addition, and to enable navigating through the results, semantic matching extracts contextual relationships between all content indexed within the system. As such, the matching can provide a list of suggested documents and articles based on the content the user is currently viewing. This list of suggestions can be organized by categories to further enhance its relevancy and positive impact on the user’s search experience. An experimental feature involves exposing facets within related content leading to something like “search guided navigation” of content.

Knowledge representation, correlation and reasoning are based on a distributed rule and event model that defines the states, consequences, as well as notifications and visualizations. This includes the definition of all relevant pieces of information needed for semantic processing throughout the different stages in the processing chain as well as definition of event sources and their semantic parameters and the definition of rules for the communication between the several components.

5.3 Search adaptation, customization & refinement

Adaptive search relates to the fact that search results are improve and are aligned to user preferences thanks to the analysis of user implicit and explicit feedback. Adaptive search techniques are related to the learning to rank paradigm [5]. It has already been well studied and used today by commercial search engines.

Recent advances in customized search are based on multi-task ranking techniques which enable a good trade-off between a user-independent search engines (high coverage but low precision) and fully customized systems (every user has different search result, leading to a small coverage but a high precision of the results).

The key innovations relate to the intelligent search methods that enable query intent discovery. This refers to structuring and interlinking an unstructured query text submitted by the user in order to improve ranking and relevance of results. It is related to the relevance feedback principle where the search interaction is composed by multiple query refinement steps. Intelligent search methods provide a quick and efficient way for users to search, refine and filter through the relevant content integrated by the system. Active learning techniques based on the value of information are used to decide which refinements could be proposed to the user [8][10]. Query intent discovery is accomplished by recognizing entities in a natural language user query and thereby deriving a machine-understandable query statement (see Figure 2).

5.4 Crowdsourcing and supervised automation in semantic matching

For semantic matching and pooling, relational learning methods are used which allow information gathered from different but related instances to be used to reason about the instance in question. The various relational representations of the content and its links (to other content, to people, etc.) are identified and provide rich information on that content. Through relational learning and reasoning, similarities and dissimilarities to other content is established so that content can be classified solely on these relation properties.

In order to utilize semantic annotations through crowdsourcing for yet unlabelled content, tensor factorization is used. It constructs a matrix of key terms with their weights from already annotated content. This matrix is factored into a term matrix and a content matrix where the latter describes clusters of related content while the term matrix contains their respective set of annotations.

Clustering is the assignment of objects into groups, called clusters, so that objects from the same cluster are more similar to each other than objects from different clusters. Clustering is a common technique of statistical data analysis. Besides the term data clustering (or just clustering), there are a number of terms with similar meanings, including cluster analysis, automated classification and numerical taxonomy.

It is important to realize that although clustering is used to find natural groupings by means of common traits in the data, by using a technique which dissects a data set into groupings, structure is implicitly imposed upon the data set. After clustering the groupings can be evaluated, named and their properties summarized. The results from clustering can be used to identify natural groupings present in the data, to perform data reduction, or to generate hypotheses for the data set that was analyzed. Finally, classification or categorization is performed in order to assign an item (e.g. cluster) to one or more categories. This decision is taken based on the values of properties of this item known as features of the content matrix.

6. LESSONS LEARNED AND POSSIBLE EXTENSIONS

6.1 Lessons learned

Initially, the matching of research profiles to funding opportunities was weak, which was one reason to employ user profiling and machine learning methods in an attempt to improve matching results. With continuing use of machine learning methods to learn from user behavior, the results for individual users improved considerably over time. This suggests that the machine learning methods indeed are able to utilize user profiling data to optimize finding and matching results. A next step would require the application of transfer learning methods in order to infer from one individual user type to a group of similar users, which would cut down the time to learn from a user and adapt to his or her preferences.

Project members have worked with research groups and SME representatives in order to define the feasibility, scope and objectives and the benefit from pooling and matching research funding opportunities into one system. Surveys were developed focusing on the type of information sources and main users involved for the selected application scenarios. A particular focus was on the relevancy of three types of information sources: in-house, partner, and external data. Specific consideration is given to the use of informal and formal information. Formal information is authored by institutional sources, such as company communications or government publications. Informal information is authored by personal sources, such as customer conversations or online discussions. Preliminary results suggest that for spotting funding opportunities, both research groups and SME representatives rely more on informal information than on official announcements or calls.

Self-administered online questionnaires will be sent to survey participants with guaranteed anonymity and no identifying information in the questionnaire. The results of the surveys are used to prioritize and reassess the different types of information sources and users and their role in the selected use cases in order to develop and refine the software. For example, if respondents rate the use of information from customers as critical, then one use case would concern the integration and analysis of related customer feedback. The use cases serve as basis for the rapid application development cycles and are reassessed and modified as needed through iterative cycles of user feedback and rapid development and evaluation.

6.2 Possible extensions

In our talks with end users, we observed a need to extend the software, which is now primarily aimed at regional development agencies, to meet requirements of SMEs.

A possible extension concerns technology intelligence and contributes to the innovation awareness and innovation capabilities of SMEs, especially those who develop innovative technologies and want to explore the opportunities for them in the market. This can be very broad, from finding partners to analyzing their position in a patent landscape.

Patent landscape analysis is becoming more important for companies since the number of patents is growing but the claims are becoming more detailed. In other words, when the patent density is increasing it is more important to be aware of the opportunities and threats in the market. Large IP driven companies are well aware of this and invest a large amount of money in continuously analyzing their IP position. They analyze their competitive strength against all other competitors over all major countries and over multiple years to catch trends. For this they need to algorithmically analyze 10k-100k or more patent documents. Their patent position is then visualized in a patent landscape. From these visualizations they might perform patent searches depending on the use case. Well known use cases are the following:

1. Patentability or novelty search where as many relevant patents (depending on technology and country coverage) are searched to determine of a technological invention is already protected by patents. The outcome can be that a new idea provides great market opportunities or that innovation in a certain direction is protected by patents of other companies. Patent and non-patent literature (where ever prior art can be described and found) need to be searched. Being able to determine innovation opportunities is essential to technical innovative SME although the large companies have the budget and expertise to do these analyses.

2. Validity search where based on a patent or a set of claims a company want to know if it can invalidate a patent and in such a way gain a competitive advantage. Invalidation is always important - especially for SME's - since it can provide a competitive edge with a small investment if one can do then analysis right and this project want to provide the tools to perform this.

3. Infringement or freedom to operate search - where one determines based on a novel idea if this is already protected by patents and then decide to proceed and invest in R&D around the new idea or one stops R&D investment around the idea since it is already well protected by patents. This is very important for SME's since they need to be able to make a difference but lack the tools of the large organization. However this project aims at providing SME's the tools to be able to be competitive on patent analysis with large patent driven companies.

4. State of the art are designed to analyze patterns and trends in large patent document sets (10k-100k or beyond) to determine the current state of the art around a technology and future trends. This is essential for the strategy of innovative companies and in this project we want to develop and provide this technology form SME's to help them to become more competitive.

7. REFERENCES

[1] Belew, R.K.. 2000. Finding Out About: *A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge Univ. Press, Cambridge.

[2] Chakrabarti, S., Dom, B. and Indyk, P. 1998. Enhanced Hypertext Categorization using Hyperlinks. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data* (Seattle, WA, USA, June 2-4, 1998). SIGMOD '98. 307-318. ACM, New York, USA, 383-392. DOI= http://doi.acm.org/10.1145/276304.276332

[3] Cheung, J. C. K. and Li, X. 2012. Sequence clustering and labeling for unsupervised query intent discovery. In *Proceedings of the 5th ACM international conference on Web search and data mining* (Seattle, WA, USA, February 8-12, 2012). WSDM '12. ACM, New York, USA, 383-392. DOI= http://doi.acm.org/10.1145/2124295.2124342

[4] Gatziu Grivas, S., Schaaf, M., Kaschesky, M. and Bouchard, G. Cloud-based Event-processing Architecture for Opinion Mining. In *Proceedings of the 2011 IEEE World Congress on Services* (Washington, DC, USA, July 4-9, 2011). SERVICES '11. IEEE Computer Society, 272 - 279 DOI= http://dx.doi.org/10.1109/SERVICES.2011.49

[5] Joachims, T. and F. Radlinski. 2005. Query Chains: Learning to Rank from Implicit Feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (Chicago, IL, USA, August 21-24, 2005). KDD '05. New York, USA, 383-392. DOI= http://dx.doi.org/10.1145/1081870.1081899

[6] Jurafsky D. and Martin J. H. *Speech and Language Processing*. Pearson Prentice Hall. Second Edition. 2009.

[7] Maynard, D., Funk, A. and Peters, W. 2009. *NLP-based support for ontology lifecycle development*. Department of Computer Science, University of Sheffield.

[8] Reichart, R., Tomanek, K., Hahn, U. and. Rappoport, A. 2008. Multi-task active learning for linguistic annotations. In *Proceedings of Annual Meeting of the Association for Computational Linguistics with the Human Language Technology Conference* (Columbus, OH, USA, June 15-20, 2008). ACL-08: HLT. Association for Computational Linguistics, 861-869.

[9] Salton, G., Wong, A., and Yang, C. S. 1974. *A vector space model for automatic indexing*. Cornell University, Ithaca, NY, USA.

[10] Settles, B. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. Univ. of Wisconsin.