LiMoSINe: Publishable summary periodic report year 2

Distribution: Public

LiMoSINe
Linguistically Motivated Semantic Aggregation Engines

FP7-ICT-2011-7-288024

Version 1.0, November 19, 2013

Project funded by the European Community under the Seventh Framework Programme for Research and Technological Development

The deliverable identification sheet is to be found on the reverse of this page.

Project ref no.	FP7-ICT-2011-7-288024
Project acronym	LiMoSINe
Project full title	Linguistically Motivated Semantic Aggregation Engines
Instrument	STREP
Thematic priority	FP7-ICT-2011-7: Language technologies
Start date / duration	November 1, 2011 / 36 months

Distribution	Public
Contractual date of delivery	1 January 2014
Actual date of delivery	16 January 2014
Deliverable number	D1.4 Second 12-monthly report
Deliverable title	Publishable Summary Y2
Type	Report
Status & version	Final draft, version 1.0
Number of pages	22
Contributing WP(s)	WP1–7
Internal reviewer(s)	Caroline van Impelen
WP/Task responsible	WP1
Other contributors	WP1–7
Author(s)	Maarten de Rijke, all partners
EC Project Officer	Pierre-Paul Sondag
Keywords	Linguistically motivated semantic aggregation engines

The partners in CoSyne are:
University of Amsterdam (UvA)
University of Glasgow (UG)
Fundacio Barcelona Media Universitat Pompeu Fabra (BM)
Universita degli Studi di Trento (UNITN)
Llorente & Cuenca Madrid SL (L&CMS)
Universidad Nacional de Educacion a Distancia (UNED)

For copies of reports, updates on project activities and other LiMoSINe-related information, contact:
The LiMoSINe Project Co-ordinator:

Prof.dr. Maarten de Rijke
University of Amsterdam
Science Park 904
1098 XH Amsterdam, The Netherlands
E-mail: derijke@uva.nl
Phone: +31 20 525 5358, Fax: +31 20 525 7490

Copies of reports and other material can also be accessed via the project's homepage:
http://www.limosine-project.eu/

Contents

1 Goal and challenges

We increasingly live our life online. Information is accumulated on a wide range of human activities, from science and facts, to personal content, opinions, and trends. Across the globe, people's knowledge, experiences and interactions effortlessly find their way to online outlets, alongside traditional edited content, ready to be shared with millions.

LiMoSINe will integrate the research activities of leading researchers across diverse topics with a view to enabling new kinds of language-based search technology. The LiMoSINe vision is to transition access to online information from a document-centric search paradigm focused on returning disconnected atomic pieces to a truly semantic aggregation paradigm. In this new paradigm, machines will understand a user's intent, discover and organize facts, identify opinions, experiences and trends, all from inherently multilingual online sources and open knowledge repositories. LiMoSINe's aggregation engines will automatically organize search results in semantically meaningful ways.

LiMoSINe has the following objectives:

1. To enable semantically structured access to multi-lingual online content;
2. To integrate deep linguistic processing in information extraction;
3. To support semantic mining where data-driven patterns are made human interpretable using the web of data;
4. To develop evaluation methods for rigorously assessing the effectiveness of semantic search and semantic aggregation in a lab-based setting;
5. To exploit its research results in three demanding multilingual use cases:
    a. open-domain community question answering,
    b. online reputation management in a professional task-based setting, and
    c. intelligent content annotation and search on a photo-sharing platform.

The components of LiMoSINe will be integrated through web services with solutions currently in place at the project's use case owners.

2 Summary of the activities carried out

2.1 Semantic scenario refinement (WP2)

Going beyond the state of the art in Natural Language Processing (NLP) brings the need for new data collections and evaluation methodologies. Approaching novel tasks in three social media scenarios Community Question Answering, Multimedia Tagging and Online Reputation Management requires appropriate development and evaluation material, as well as a suitable framework for benchmarking activities. Most of the work done in this year in WP2 has been focused precisely on data set creation and defining appropriate evaluation methodology.

One of the main WP2 activities has been the organization of RepLab 2013, an evaluation campaign for Online Reputation Monitoring technologies coordinated by UNED, UvA and Llorente & Cuenca.[1] As many as 45 research groups from 22 different countries signed up, showing a growing interest in this competitive evaluation exercise. The results allowed us to assess the current capabilities and technical limitations of the

state-of-the-art systems designed to tackle tasks related to Online Reputation Monitoring (ORM).

The dataset created for the purposes of RepLab by the volume of manual annotations constitutes a unique resource for developing and testing systems and system components for ORM in Twitter. Besides, it enables a better understanding of what is behind the reputational analysis performed by the experts.

From a methodological point of view, the organisation of RepLab 2013 motivated work on general evaluation measures applicable not only to the problems related to ORM, but also to a number of document organisation tasks. This research will enable standardising the evaluation process and providing more robust and informative metrics.

Significant effort has been dedicated by the University of Trento (UNITN) to build SenTube, a corpus for sentiment analysis. Although the data collection is not finished yet, the manually annotated English comments on YouTube videos about tablets and cars have already been used in experiments on comment type and sentiment classification.

There are two new image collections based on Flickr created by the University of Glasgow (UG) for Multimedia Tagging tasks: Automatic Image Annotation and Photo Tag Recommendation. Another data set produced by UG in this period is an annotated collection of tweets for studying event detection in Twitter.

Finally, in order to provide the partners with a common data collection for developing and evaluating algorithms and tools produced within the Consortium, we have started building a LiMoSINe *Common Corpus*. Composed by news articles from Google News, tweets and YouTube videos with accompanying user comments, this multilingual resource, will constitute a valuable testing ground and a source for future datasets for Information Access on social media.

2.2 Information extraction through deep linguistic analysis (WP3)

The growing demand of online users for new functionalities of search engines, which is going far beyond simple keyword-based search, brings us deeply into the natural language processing territory. There is a growing need for a *deeper semantic interpretation of data,* requesting a semantic aggregation model. At the core of such a linguistically motivated semantic aggregation model is a representation that is able to capture various objects of interest (e.g., entities, opinions) and their relationships. One of the goals of the project is to develop such a *semantic model extractor*, which is an automatically extracted semantic representation of various levels of linguistic annotation. The semantic model is based on disambiguated entities, relations between them, subjective expressions, opinion holders and relations among these pieces of semantic information.

During the first year of LiMoSINe, an initial prototype of the semantic model extractor was developed using the Apache UIMA (Unstructured Information Management Architecture) framework and deployed as web service. An overview of the overall pipeline of the semantic model extractor is shown in Figure 1.

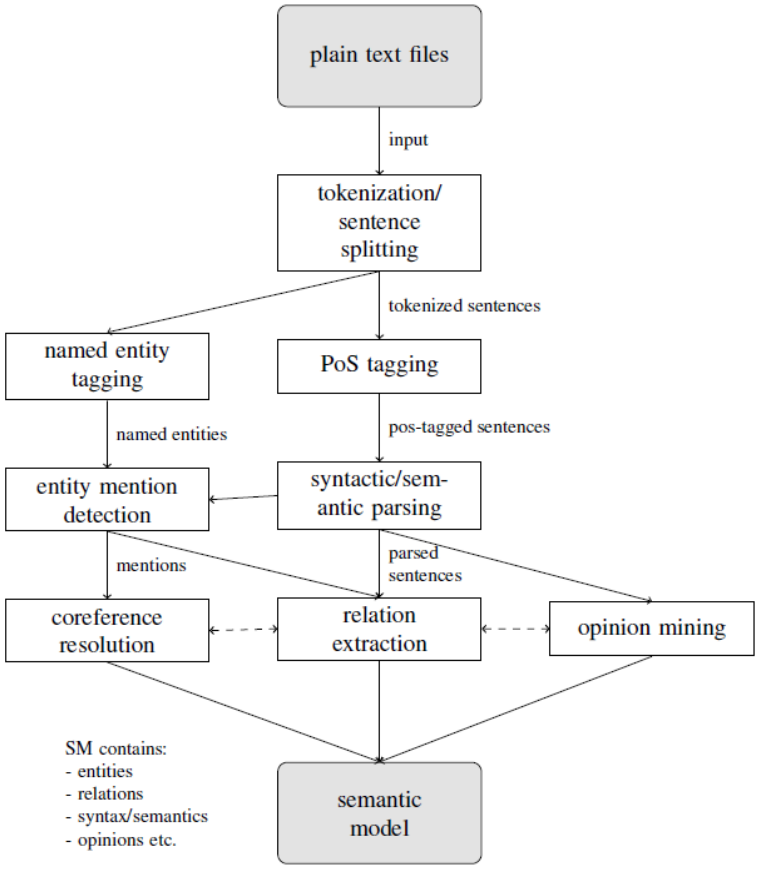


Figure 1: Overview of the UIMA pipeline.

During the second year of the project, we focused on multi-linguality. For the second release of the prototype, we created a semantic model extractor for Italian, comprising (a) tokenization and sentence splitting, (b) named entity recognition, (c) parsing, (d) mention detection, (e) relation extraction and (f) coreference resolution. We have also continued our work on the Semantic Model for English, providing different models for relation extraction and incorporating an entity-linking module. We are currently in process of building a Semantic Model for Spanish to be finished in the nearest future.

The pipeline, along with its individual components, has been evaluated on various datasets and has been used to create submissions to several shared tasks, including *SEM-2012 (Semantic Textual Similarity) and CoNLL-2012 (Coreference).

Our goal for the final year of the project is two-fold: (a) we plan do elaborate on specific components of our prototype, improving their performance and extending our evaluation experiments to assess the impact of the deep integrated semantic representation on various information retrieval tasks and (b) we want to build upon our multilingual pipelines, identifying cross-lingual mappings between individual semantic models.

2.3 Semantic mining (WP4)

The main objective of this work package is to mine meaningful patterns and properties of information objects relevant to the scenarios considered by LiMoSINe. In particular, new dynamic semantic search tasks are defined and the web of data is used to provide

descriptions of patterns mined from text streams: around entities, events, online user behavior, and images. Information, text-based or semantic, pertaining to a single information object is aggregated to prepare for semantic aggregation search tasks.

In the first year, we mainly focused on semantic aggregation (in the form of entity modeling and entity linking) and on modeling automatic media annotation. The objective of semantic aggregation is to create a large knowledge base regarding information available in a set of documents, such as webpages, blog posts, tweets, or community question answering websites. We aim at identifying entities in the documents and characterizing the connections among them. This knowledge will be used for the task of composite retrieval, where a user can query about an entity, or a small set of entities, and the system will compose a concise answer containing what is known about those entities. For modeling automatic media annotation, we have improved the performance of an existing state-of-the-art annotation model by exploiting tag co-occurrence and temporality. The first approach places preference to tags that frequently co-occur with suggestions made from an SVM. We were able to achieve statistically significant improvements to annotation accuracy on a collated Flickr collection.

The deliverables of WP4 in year 1 consisted of baseline versions of systems to address the tasks mentioned above. These include a semantic mining module, that offers facilities for automatically linking textual documents to structured knowledge bases in various languages; results were published at WSDM 2012. Additionally, a media annotation module for the automatic annotation of images using contextual information such as tag co-occurrence was produced as well as a semantic module for obtaining the historical interest and development around concepts; this module was released as a public API (http://www.opengeist.org); see Figure 2.

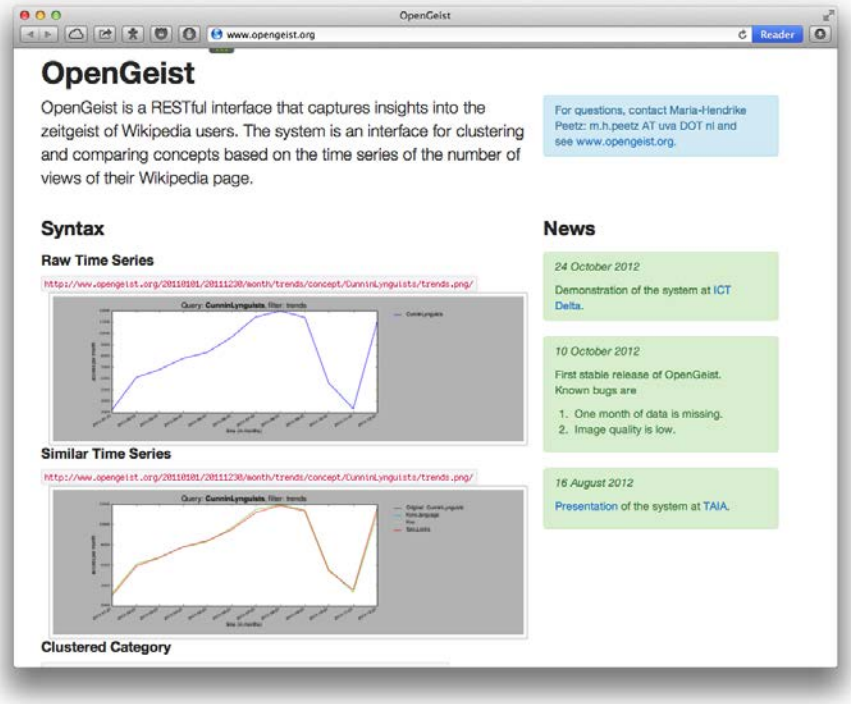


Figure 2. Opengeist home page.

In the second year, we extended our work on semantic aggregation and automatic media annotation and released a revised version of our online behavior predictor. The revised version of our semantic aggregation module allows entity linking in document streams of short text, which is useful in second screen applications, e.g., contextualizing running subtitles in TV programs. Additionally, we challenged our entity linking approach in domains where Wikipedia might not offer enough coverage. We turned into the music domain and developed methods for linking music-related tweets to songs and artists. We confirmed that Wikipedia's coverage in this fast paced domain is limited and turned to Musicbrainz and YouTube as additional knowledge bases. The use of these additional knowledge bases increased the entity linking effectiveness of our method. This exercise led to a new demonstrator, which we call Streamwatchr. Our revised media annotation module takes into account and combines additional signals (e.g., temporal, geographical, and visual) achieving better annotation effectiveness over simple co-occurrence of tags. For our online behavior predictor we developed a learning to rank framework that monitors and tracks user behavior as manifested from clicks or other similar signals. This framework is ready to be plugged into systems that work in real-time settings and learning and adaptation need to happen online.

2.4 Access and recommendation (WP5)

The main objective of this work package is to investigate novel techniques for effective access and recommendation of information by mining a set of disparate textual and semantic resources. The research work concentrates on data streams like Flickr (multimedia) and Twitter (Social streams). Important results of second year research include: A Flickr data set of 2M images and tags has been crawled and a research test-bed is created (ACM MMM 2013). A tweet event detection test-bed is created with 120 million tweets and more than 506 events annotated (ACM CIKM 2013). An fMRI study analyzing information retrieval and brain reactions is published in the ECIR 2013 conference and was given the best-paper award. The work has been disseminated in a number of prestigious venues like ECIR 2013, ACM SIGIR 2013, ACM MMM 2013, CIKM 2013 and WWW 2013.

A module for cross-lingual information access was developed. The goal of this module is to offer novel techniques for cross-language information retrieval. It combines results from traditional cross-lingual information access methods with a newly developed method, which is based on semantic information, i.e., recognizing entities and perform a matching between queries and documents based on this form of semantic enrichment. We developed a data set for conducting experiments on tag recommendation and will be presented at the ACM MMM 2013. We proposed a photo tag recommendation system, which automatically extracts semantics from visual and meta-data features, to compliment existing tags. Compared to standard content/tag-based models, these automatic tags provide a richer description of the image and especially improve performance in the case of the cold start problem. In addition, we studied the role of temporality in tag recommendation. The results are presented at ACM SIGIR 2013 conference. In addition, a new technique for detecting events from social streams (Twitter) is studied and a paper is submitted to WWW 2013. An evaluation test-bed is created and is presented at the ACM CIKM 2013 conference. The data set is made available under Twitter Terms and Conditions for the community.

The paradigm of composite retrieval consists of extracting from a collection of information items, a set of diverse item bundles that together form the best possible answer to a user's complex information need. We proposed and experimented with a

number of Bundle formation techniques and a research paper is submitted to the WWW 2014 conference. The implementations of some these algorithms constitute the nucleus of the composite ranking module that will be integrated in the entity-Driven Exploratory and sErendipitous Search SystEm (DEESSE Demonstrator) system, aimed to support a serendipitous exploration of complex data extracted from different social media. There has been growing interest in building aggregated search systems where information from a variety of sources (so-called verticals) is retrieved and aggregated into one single interface. We are developing a framework for evaluating as well as optimizing aggregated search systems in terms of relevance and diversity of information. In the second year, various aspects of our evaluation framework is studied and presented at the ACM CIKM 2013, and WWW 2013 conferences.

Our research activities on context modeling and personalization focused on searchers and their behavior. The starting point is to understand the nature of relevance from a neuroscience perspective. Therefore, we continue our investigations using techniques to comprehend how the human brain functions when a relevance judgment is issued. The objective is to understand the brain functional behavior and how it is associated with relevance judgment. We investigated the connection between relevance and brain activity using functional Magnetic Resonance Imaging (fMRI). In our study, we measured the brain activity of eighteen participants while they performed four topical relevance assessment tasks on relevant and non-relevant images. The results revealed three brain regions in frontal, parietal and temporal cortex where brain activity differed between processing relevant and non-relevant documents—an important step in unraveling the nature of relevance and building better user modeling techniques. The result of this work published at the ECIR 2013 conference was given the best-paper award. A further study is designed with a sophisticated experimental design and 29 participants were scanned while they perform video search. The results are being analysed. We also studied the user context, from emotional, physiological, and behavioral perspectives in an information seeking context. Emotional reactions are captured with a web cam and analysed. Similarly physiological reactions and behavioural data analysed. The results include emotions and other features help in identifying relevance and two research papers are published at the ACM SIGIR 2013 and WWW 2013 conferences.

3 Demonstrators

Many of the semantic aggregation technologies built as part of LiMoSINe will be integrated into three demonstrators. Important achievements are the full specification of the three demonstrators in terms of scenarios and test-beds that allow us to showcase the new technologies developed as part to the projects, and the first prototyping of these demonstrators.

3.1 Photo tag recommendation

Despite decades of research the task of automatically “understanding” an image, identifying what types of objects it contains is still not possible for modern day computers. Most advancement in this area has been made in focused applications, e.g. in face recognition, landmarks detection, scene classification. However, no system comes close to a general-purpose image labeller, covering anything from trees, to people, to buildings, to insects. Standard image search systems therefore rely either on textual clues about an image content or on metadata, such as the image date, its size or the person who uploaded it. Such information can sometimes be inferred, e.g., from anchor text of hyperlinks pointing to images, but more and more users also provide such information directly by tagging their photos. For this reason, image hosting websites

such as Flickr encourage their users to provide tags for their photos in order to improve their “searchability.” In order to reduce the cognitive and physical effort for the user, tag recommendation methods have been introduced in many publications.

The purpose of tag recommenders is to assist users in choosing descriptive, relevant tags based on some available knowledge. Photo tag recommendation can be viewed as a retrieval problem with a composite query, which may include various properties of the photo e.g. tags already specified by the user etc. Special-purpose features mined through co-occurrences add a semantic layer to our initial knowledge about the photo. This is important especially in a case of a "cold start", when user has not provided any initial tags. Our system is built on a dataset of 1M user images, which were crawled from Flickr. A user first uploads a photo, adds 0 or more tags, before a number of features are extracted (e.g. textual, visual, contextual and user features). Based on these features, a list of 10 tag suggestions are computed by the recommendation model and are offered to the user.

In our approach, we attempt to improve these suggestions by considering additional textual evidence from Twitter. Given an image, we attempt to find a number of related tweets from Twitter, based on the image's location, time taken, user and textual content. We hypothesise that there exist a number of images taken at social events (e.g. concerts) which will have a textual commentary on Twitter which can be exploited for tag recommendation purposes. Given a list of related tweets, we then offer the top 10 most significant terms as tag suggestions to the user.

3.2 Entity-driven exploratory and serendipitous search system

The users of our entity-Driven Exploratory and sErendipitous Search SystEm (DEESSE) are the general web users. Starting from some entity (a person, place, event, etc.) as a query, DEESSE allows the user to explore entities, potentially finding other interesting and serendipitous information. These entities are part of a network built from two sources user-generated content. Our network extraction methodology has evolved since last year: we changed the similarity measure used to connect entities from simple co-occurrence in documents to cosine similarity of the whole sets of documents where two entities appear, and we refined our random-walk based retrieval algorithm. In addition, in this year, extensive evaluation of our approach was carried out, leading to many insights about the performance of DEESSE.

Datasets

We consider two user-generated content sources upon which we extract entities: Yahoo Answers and Wikipedia. Using the two sources allowed us to experiment with two very different types of user-generated content: Yahoo Answers (YA) consisting of 67,336,144 questions and 261,770,047 answers, and Wikipedia (WP) consisting of 3,795,865 articles.

Entity Networks

We call entity any concept that appears as a Wikipedia page. We describe the extraction process of entities from our text corpora in the previous annual report. This process results in 896,799 distinct entities from YA, and 1,754,069 from WP. Then starting from the set of entities extracted from each dataset, we construct an entity network by using a content-based similarity measure to create arcs between entities, which is based on a textual representation of each entity by putting together all of the documents where the entity appears.

Retrieval Algorithm

Random-walk based algorithms such as Personalized PageRank have been applied in many recommendation problems. Our algorithm for extracting from a network the top *n* entities that are most related to a query entity, is inspired by the above research line.

Metadata

We extract from our dataset information regarding quality, sentiment, and topical categories. The metadata features are first collected at document/sentence level, and then aggregated to derive scores for all the entities in a dataset.

Yahoo answers:

Wikipedia:

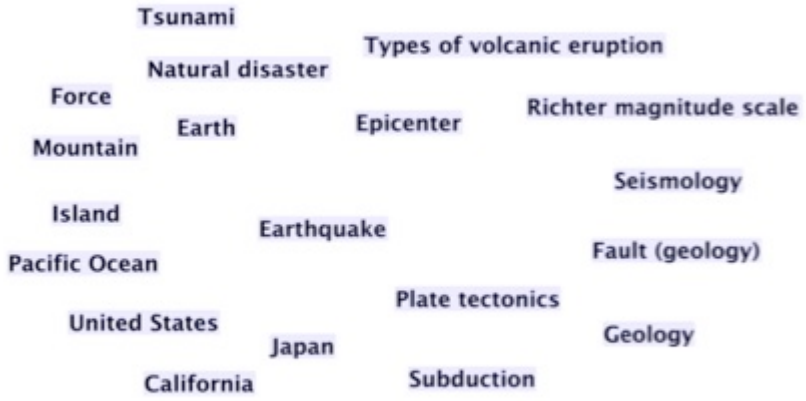


Figure 3. Outcome of a the query network for the entity “earthquake” for two datasets.

Figure 3 depicts the outcome of the query network for the entity “earthquake” for the two datasets. For both data sets, we obtain typical examples of entities related to the query.

However, we can note that for Yahoo Answers dataset, we have entities related to praying and god.

Testbed

We tested the performance of our system using a set of test queries. These 50 queries, including people, places, websites, etc. were the most queried in 2010 and 2011, as listed by Google Zeitgeist.

Retrieval Performance.

For each query, we retrieved related entities from the YA and WP entity networks. We then used CrowdFlower.com - a virtual marketplace for micro-tasks - to assess the relevance of the top 5 results retrieved for the queries in our testbed. The retrieval algorithm performs somewhat similarly on the two datasets. When we examine the ranking performance of our algorithms by comparing the Mean Average Precision scores - 0.716 (WP) and 0.762 (YA) - to precision, we see an improvement in scores, indicating that the relevant entities tend to be shown at the top of the rankings.

Although the two datasets have comparable performance, the overlap between the results is very small - an average of 0.6 entities (that is under one result) in common in the top 5. This suggests that combining the results would improve recall, and perhaps introduce more diversity. To verify this hypothesis, we aggregate the two rankings extracted from the two datasets, achieving a Mean Average Precision of 0.782, which improves the performance on the individual datasets.

Serendipity and Interestingness

We conduct an extensive study to understand what these two different sources of user-generated content can provide to serendipitous search. We consider a basic scenario in which, for our 50 queries, we compare the results extracted from the two entity networks (YA and WP). Then, we constrain the retrieval in the dimensions of sentimentality, quality and topical category by filtering the results of the original retrieval so as to select the top results that satisfy a constraint. The constraints are: one shared category with the input query, high/low sentimentality, and high/low readability.

To compare our exploratory-search setups in terms of serendipity, we adopt a metric designed to capture two essential aspects of serendipity, unexpectedness and relevance. For relevance, we use the editorial judgments collected through the annotation task described before. Unexpectedness is measured by comparison with benchmarks that produce expected recommendations. We use four baseline generators of obvious recommendations, considering the entities that occur in the top 5 search results provided by two major commercial search engines, or in the related query suggestions. Our experimental setups achieve higher serendipity than the baselines, especially when considering the topic-constrained or the unconstrained cases. The low-sentimentality and low-readability setups perform considerably worse, due to the fact that these constraints seriously hurt relevance, as mentioned before. YA always outperforms WP.

We next evaluate other more subjective aspects of serendipitous search by performing another set of crowd-sourced evaluations. Besides being relevant to the query, the results must be interesting enough to the user to catch his or her attention, and to encourage further exploration. To make sure we separate intrinsic interestingness of entities from the extent to which a user interested in a search query is interested in a

presented result, we ask labelers to consider both questions. Furthermore, we attempt to measure the value of the results by asking whether the result allowed one to learn something new about the query entity.

We take four evaluation dimensions into account: relevance, interestingness for the query, interestingness to the user, and learn something new about the query. Due to the highly subjective nature of these dimensions, we compare the results of our various experimental setups to each other instead of attempting to assign an intrinsic interestingness value to each result. We perform pairwise comparisons between all of the result pairs and build a reference result ranking for each dimension. The difference between the result ranking in each run and this reference ranking can then be used to gauge the difference between the various runs. We used CrowdFlower.com. The query and results were shown along with their Wikipedia pages, and four questions were asked:

1. Which result is more relevant to the query?
2. If someone were interested in the query, would they also be interested in these results?
3. Even if you are not interested in the query, are these results interesting to you personally?
4. Would you learn anything new about the query?

For all questions, YA produces results ordered similarly to the reference rank. In fact, for the general, unconstrained runs YA produces better rankings than WP for all four questions. The correspondence is more pronounced for question 3, which concerns personal interest in the entity. The difference is especially striking, considering a nearly even share of results from both datasets in the top 5 entities of each reference ranking.

Constraining search results using topical category improves performance only for YA. Adding a high-sentimentality constraint also boosts performance for WP, but the same is not true for YA, where the lack of editorial oversight allows for low-quality highly-emotional posts. Whereas the low-sentimentality constraint hurts performance for the group sports for WP, the opposite is true for YA. Possibly, the already emotionally-intense sports discussions in YA benefit from a selection of less intense documents.

Support for Spanish data

We started development on the multilingual aspect by adding Spanish as a second language. We investigated to which extent the workflow that was designed for the English data can be transferred to include another language. We filtered the Yahoo Answers data for Spanish documents, yielding 17,133,155 questions and 66,380,738 answers. As entities should be extracted from these documents, we are currently analyzing the capabilities of entity extraction toolkits, including an internal Hadoop-based grid implementation and a web-based service from the University of Amsterdam. Given the amount of data to process, we are investigating the achievable throughput for each of the tools. In parallel, we are building a ground truth of manually annotated entities, which we will use to evaluate the accuracy of the above tools with respect to entity extraction in Spanish.

Composite Ranking

We started improving the initial ranking module by adding on top of the random-walk based approach described above two simple instantiations of the composite-ranking module. Given an initial query provided by the user, this initial module constructs a

composite solution that attempts to cover all the different aspects of sub-topics of the query, by retrieving a specific bundle of information items (entities) for each topic. In other words, we use the possible topics of the input query to define intra-bundle compatibility and inter-bundle diversification constraints.

The main method we adopt to discover the possible sub-topics of the input query is an approach based on query-log analysis, which interprets the specializations of a query as the possible subtopics or intents of that query. We started implementing this approach by processing a large log sample of the Yahoo search engine, spanning the same two years (2010–2011) covered by the Yahoo Answers dataset. We used a query-flow graph model to segment the log into logical sessions and to extract query specializations. Using the same entity-extraction tools that we applied to build the entity networks, we mapped the query specializations onto entities. We are now working on an improved version of the retrieval algorithm, which will consist on a random walk with restart to the entities representing the subtopics of the input query (with preference proportional to the importance of each subtopic).

As a simple alternative to build a topical bundling of entities, we have also designed an approach that exploits Yahoo Answers categories. We begin by examining the result set for an input query, and the categories associated with each entity in this result set. Each entity has up to three categories (the three most frequent categories associated with the questions and answers where the entity appears). We try to find the common categories between these entities, and group them accordingly. We then try bundling them with these categories. Thus, a bundle is associated with a category or a set of categories.

The algorithm can be parameterized in order to achieve a desirable distribution of entities between the bundles, controlling (i) the number of bundles and (ii) the number of entities in these bundles that we want to get.

3.3 Online reputation management

The rise of social media has brought serious concerns to companies, organizations, and public figures on what is said about them online; one of the main reasons is that negative mentions may affect businesses and careers. Monitoring online reputation has therefore become a necessity. An online reputation analyst typically has to perform at least the following tasks: given a stream of texts containing the name of a company as input, (i) *filtering* out tweets that are not related to the entity of interest, (ii) determining the *polarity* (positive, neutral or negative) of the related tweets, (iii) clustering the strongly related tweets in *topics*, and (iv) assigning a relative priority to the topics, in terms of whether it may damage the reputation of the entity.

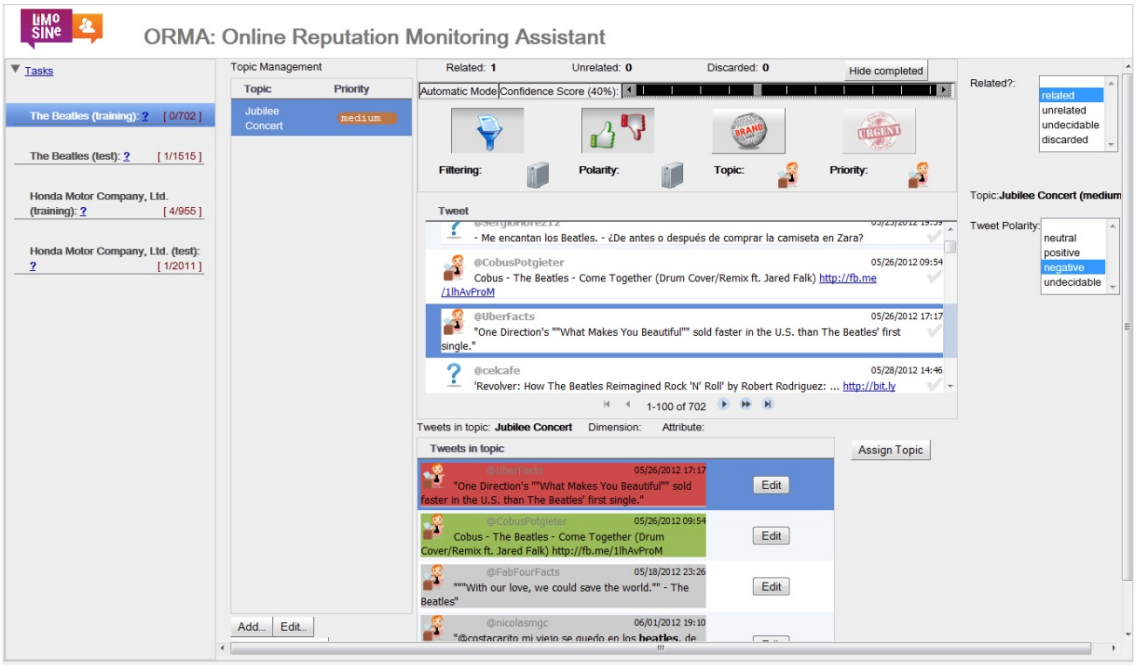


Figure 4. User interface for the Online Reputation Monitoring Assistant.

The Online Reputation Monitoring Assistant (ORMA) demonstrator aims to assist the daily work of reputation experts, helping them to process the data more efficiently. The interface helps the expert to understand the content being analyzed. To reduce effort, the system also proposes different automatic labels for each input data, with a confidence score indicating how trustable these labels are. These automatic annotations can be manually changed by the reputational expert. The input data of the demo are tweets in both English and Spanish.

An earlier version of the ORMA annotation application (which did not include the option to automatically process the data) has been tested by thirteen experts during the preparation of the RepLab 2013 test collection. Over half a million annotations for 61 different entities and more than 142.000 tweets were performed for a total workload of 20 person-month. During the exercise, the application was extensively tested for robustness and user friendliness. In particular, interaction design was significantly enhanced by many GUI changes suggested by the annotators.

3.4 ThemeStreams

Over the past couple of years, politics and politicians have discovered social media as important means for communicating with voters and for influencing public opinion. Keeping track of the many discussion forums and other outlets is no trivial matter. What themes are being discussed? Who introduced a theme? Who ``owns'' it? Typical politically relevant themes include: the economy, healthcare, defence, foreign policy. According to a leading communication agency, during recent national elections in The Netherlands discussions revolved around approx. 500 issues, with differing levels and patterns of attention.

The participants of political discussions can often be mapped to a select number of so-called influencer groups. Specifically, one can identify the following four groups. First, there are those who currently actively have an (important) position within the governing body, the politicians. Second, there are those who lobby for (specific) important issues,

the lobbyists. Third, there are journalists who specialise in politics as well as other high profile media influencers such as television stars or columnists. Fourth and finally, all other people taking part in political discussions we group together as the rest: the public.

In this demonstrator we describe ThemeStreams: an interactive visualisation aimed at giving insight into the ownership and dynamics of themes being discussed, thereby enabling users to answer questions such as *Who put this issue on the map?* ThemeStreams allows users to explore streams of tweets, either from a fixed set of predefined themes or through a search box. It uses stream graphs to indicated how influencer groups discuss a theme, thereby depicting the ``aliveness'' and ownership of a topic. Our visualisation indicates when somebody said something, which influencer group this person belongs to, and it takes into account how many people react to a statement to estimate the ``size'' and ``lifetime'' of a statement. ThemeStreams can be accessed at http://themestreams.xtas.net/.

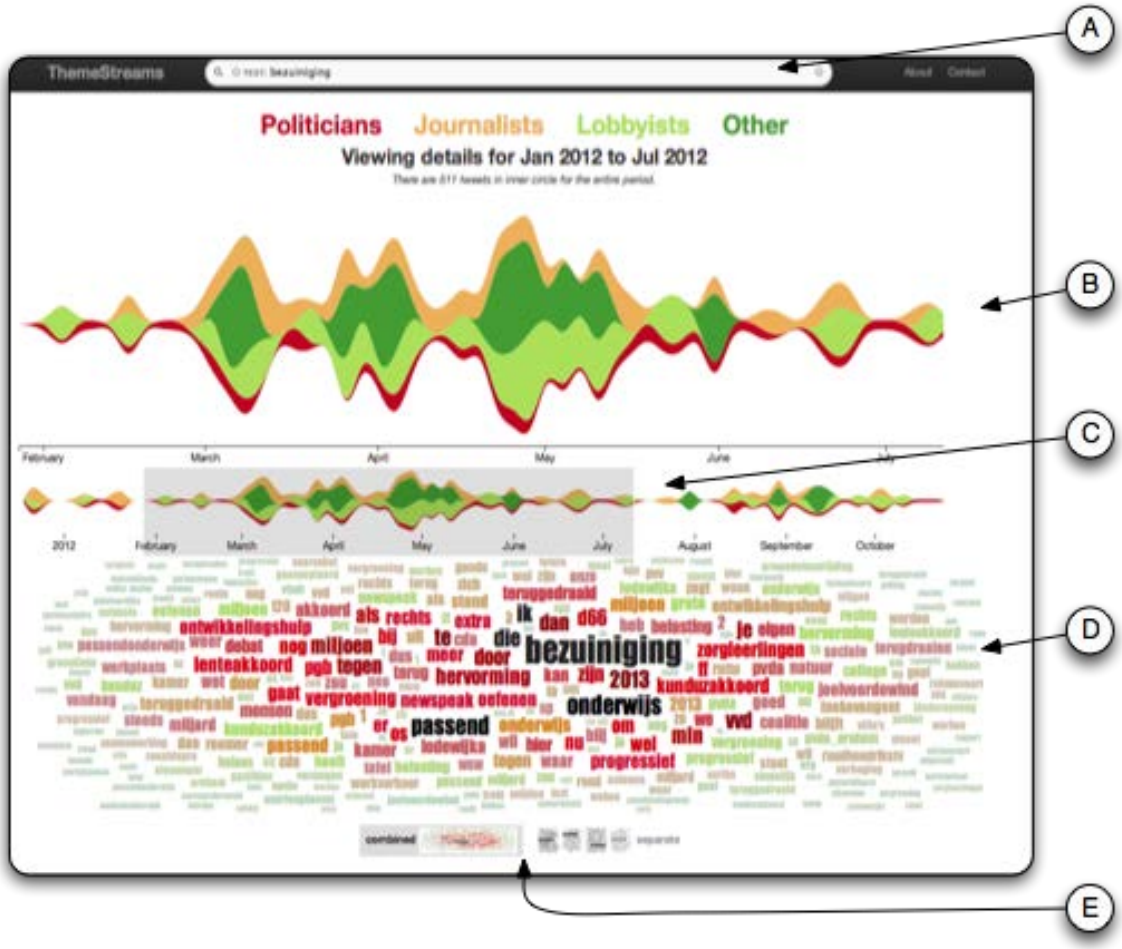


Figure 5. The interface of ThemeStreams. A user enters a topic they are interested in (A), a stack of streams coloured according to the respective user group is shown over time (B). Selecting a time period (C), one can see the most popular terms discussed by each user group using a term cloud (D). Two types of term clouds are available (E): one which is generated over all user groups, and one generated per user group.

3.5 Streamwatchr

Social media is changing the way we consume music. Online music services such as iTunes, Spotify, last.fm, and YouTube, enable us to access music from everywhere, anytime, and share our playlists with the world in the form of tweets, or status updates. People tweeting the tracks they are currently listening to generate more than half a million tweets per day. This offers us insights into people's music listening behavior at world scale. In this demonstrator, we present methods for converting free, unedited text to structured data, which we, then, use to analyze the world's music listening behavior. We choose to focus on Twitter because of their large user-base, and easy access to their content.

There are two major challenges in mining social media for studying music listening behavior, except the sheer volume of incoming data: (i) how to identify music related content, and (ii) how to deal with the semi-structured, unedited nature of user generated content. For the first challenge, we use four popular music hashtags on Twitter for identifying tweets potentially related to music: #iTunes, #nowplaying, and its shorthand, #np, and #spotify. Tweets tagged with #iTunes and #spotify are automatically generated by the respective software music players, while those tagged with #nowplaying are not associated with a particular source and may contain additional information to the track, e.g., lyrics, or experiences.

For the second challenge, we use a set of regular expressions to generate a candidate set of artists and songs, which we curate using the Musicbrainz.org database (an open music knowledge base). Our difference with previous work is that we use Youtube search for increazing the recall of our method, and develop tailored webpage extractors for the #iTunes and #spotify tagged tweets; we provide a comparison of these methods in the next section. Another dimension to this challenge is to identify the geolocation of a tweet. We use Geonames.org, an open geo database, for mapping a tweet's coordinates (or extract the twitterer's coordinates from their profile description, or location) to a geolocation.

In this demonstrator we describe Streamwatchr: a real-time system for analyzing music listening behavior at world scale. Streamwatchr aims at (i) mapping unstructured, user generated content to structured data in real-time, and (ii) providing up-to-date visualizations of what the world is listening to, what songs and artists are trending, and what will be the next big music hit. Streamwatchr can be accessed at: http://streamwatchr.com/.

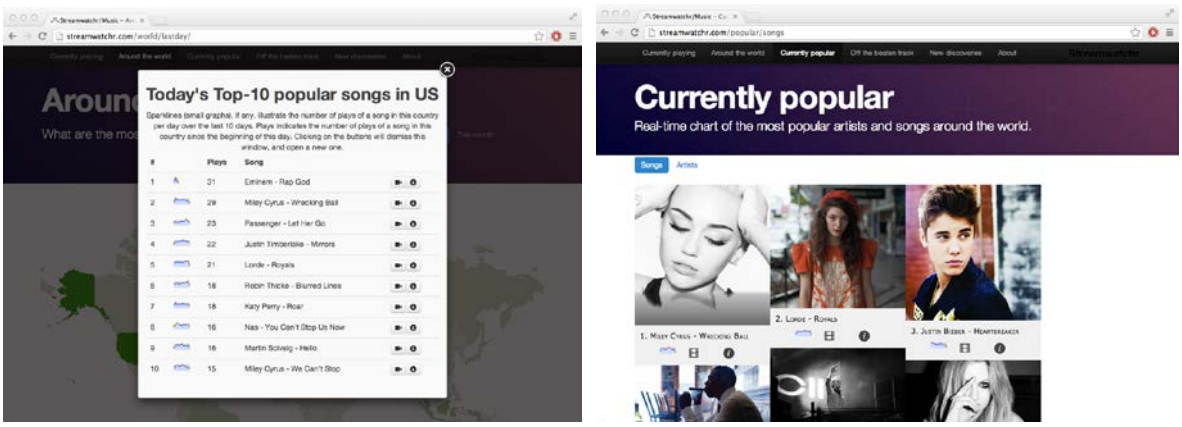


Figure 6. The interface of Streamwatchr. After extracting songs and artists from music-related tweets, Streamwatchr compiles top-10 charts per country (left), or aggregates the statistics for providing the top-10 most popular songs in the world, in real-time (right).

4 Dissemination and exploitation

The second dissemination and exploitation plan for the LiMoSINe project provides an overview of the planned dissemination and exploitation activities, achievements and results for the second year of the project.

During this second year, we have attended scientific events and a meeting organized by the Reputation Institute with relevant industry contacts. We have contacted over one hundred journalists and bloggers from The Netherlands, The United Kingdom, Italy, Spain and Belgium. Dissemination material for this period includes an infographic to promote RepLab 2013. Also includes an outline of the exploitation plan and a market study of online monitoring tools.

The main objectives of this second year were:

- **To get qualified opinion**: to engage qualified audiences through our messages (opinion leaders: journalists, authorities and bloggers). Our messages will target specific audiences.

- **To get international media coverage**: to obtain qualified media coverage, we have identified the most important and specialized blogs and newspapers in each country (The Netherlands, United Kingdom, Italy, Spain and Belgium).

4.1 Dissemination activities

Presentation of research results

This year, UNED, UvA, and LLORENTE&CUENCA have organised the second campaign, RepLab 2013. One of its main outcomes is its dataset, the largest existing test collection for research problems related to online reputation monitoring. The results of Replab 2013 were presented in September at a workshop at CLEF 2013.

Participation at relevant industry conference

In June, we attended ‘The Reputation Management Journey’, the International conference on corporate reputation, brand identity and competitiveness organised by Reputation Institute, the world's leading reputation management consultancy.

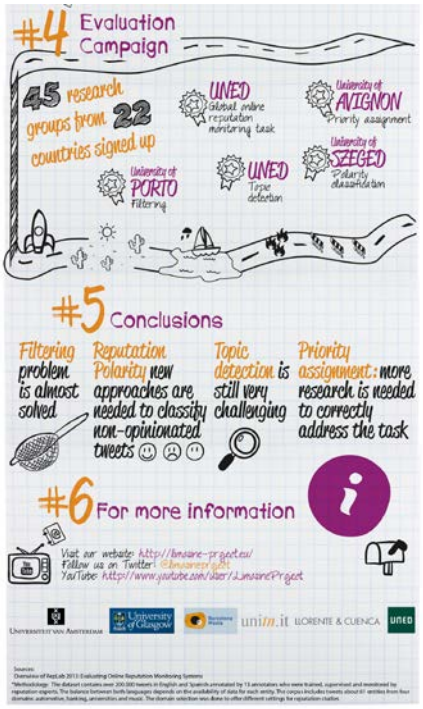
The main European companies attended this meeting. These enterprises were: BBVA, Aqualogy, Procter & Gamble, PwC, Caixabank, Correos, Carnival Cruises Line, IE Business School, DKV, Gas Natural Fenosa, Global Alliance, Iberdrola, Interbrand, Itaú Unibanco, Nestlé, LLORENTE & CUENCA, Ogilvy & Mather, Repsol Pirelli, Vestas Wind, C&A Brasil, Diageo, Huawei, ING Direct and Banco Santander.

The main objective of our participation in this meeting was to bring together the scientific community and the relevant industry. Our participation was very important because we tried to disseminate the best ideas of our research to the main European companies, so they can know us better and eventually have all the trust in the ability of our products to make them succeed.

These European companies are ones of the most important stakeholders for the project and we are constantly working on creating the best products to fit their necessities and wishes.

4.2 Dissemination materials

RepLab 2013 infographic: we have done an infographic to announce and promote de Replab's 2013 results. The infographic presents the evaluation campaign, providing a definition of the full task and its components, describing the dataset, giving the basic data on the participation and results, and drawing some conclusions.



Social media presence:

a. **Twitter**: active since project month 11.

b. **YouTube**: online since project month 20.

c. **Slideshare**: online since project month 23.

4.3 Exploitation plan

The second Exploitation Plan, apart from the market analysis and business opportunities, includes exploitation plans by the different consortium partners of LiMoSINe outlining their views and ideas on potential use of the tools. Detailed application and implementation results by each of the partners will be provided in the third and last dissemination and exploitation LiMoSINe report.

The online reputation monitoring tool

The main objective is to identify trends in the online reputation monitoring tools market, focusing on the European and American market.

The current report provides an overall guideline of the exploitation strategy, focusing on the market analysis (monitoring tools). It will detail more on dissemination and exploitation aspects.

The exploitation strategy includes three principal components:

- Market analysis: benchmarking of monitoring tools.
- Business plan: detect business opportunities.
- Market survey.

The market analysis has been designed to determine the state of the art in online monitoring tools. It analyses the main advantages and disadvantages of the commercial systems available on the market and assesses their appropriateness for the daily analysis of online conversations about companies.

The following is a list of current trends in the sector, based on the analysis of monitoring tools. It presents the limitations and the benefits of use a monitoring tool. We can conclude that:

- There is no tool that performs searches in all the Internet sources.
- Semantic disambiguation is still a work in progress.
- No real-time updates of a company, brand or product mentions.
- The classification of the mentions on topics is not a common function in monitoring tools.
- The automatic analysis of polarity is not reliable.
- Human feedback is essential to make good use of the monitoring tools.