D5.1

Text mining and linking modules – v1

v 1.0 / 2015-01-13

Adrian Popescu (CEA), Etienne Gadeski (CEA), Symeon Papadopoulos (CERTH), Romaric Besançon (CEA)

The current deliverable is a technical report accompanying the first version of the USEMP text mining and linking modules developed during the first iteration of the project. The primary objective of these modules is to process the text content that is associated with the users of Online Social Networking (OSN) services, .e.g. their posts and comments, the content of the articles they like, etc. in order to extract personal information cues that could be used for user profiling.
In particular, this deliverable documents the underlying principles and methodologies of the developed modules, the exposed functionality, the respective implementation details, and the conducted evaluation experiments. In addition, it highlights the importance of each module for the project use cases and the multi-disciplinary issues arising from their deployment. During the first iteration of the project, work focused on two modules, namely multilingual text similarity and text based location recognition.

Project acronym	USEMP
Full title	User Empowerment for Enhanced Online Presence Management
Grant agreement number	611596
Funding scheme	Specific Targeted Research Project (STREP)
Work program topic	Objective ICT-2013.1.7 Future Internet Research Experimentation
Project start date	2013-10-01
Project Duration	36 months

Workpackage	WP5
Deliverable lead org.	CEA
Deliverable type	Prototype
Authors	Adrian Popescu (CEA) Etienne Gadeski (CEA) Symeon Papadopoulos (CERTH) Romaric Besançon (CEA)
Reviewers	Tom Seymoens (IMINDS) Theodoros Michalaereas (VELTI)
Version	1.0
Status	Final
Dissemination level	RE: Restricted Group
Due date	2014-12-31
Delivery date	2015-01-13

Version	Changes
0.1	Initial input by CEA
0.2	Input from CERTH
0.3	Refinements from CEA
0.4	Refinements from CERTH
1.0	Refinements after internal reviews

Table of Contents

1. Introduction

This deliverable provides a description of the USEMP text similarity and location prediction modules implemented during the first iteration of the project. The introduction first gives an overview of the role of text mining in USEMP, of the research methodology and of multidisciplinary interactions within the project.

The main objectives of the deliverable are:

a) to clarify the usage of text mining modules in the USEMP framework;
b) to detail the research approaches adopted, including implementation details,
c) to present an evaluation of text mining modules;
d) to detail how these modules are interfacing with other modules in the USEMP system.

1.1. Text mining and linking in USEMP

The main objective of text mining and linking modules in USEMP is to endow the system with the capability to **conduct inferences about OSN users' interests and traits based on the content of the texts** they share and interact with. Inferences are extracted for individual texts, but are subsequently exploited in other parts of the project, as follows:

- Direct usage of text-based inferences in the system implemented as part of WP7;
- Combination with inferences based on visual content processing in D5.3, followed by integration in the system;
- Fusion with behavioral cues in the privacy scoring framework described in D6.1, followed by integration in the system;

The types of information that can be inferred by processing users' texts include a wide variety of personal information such as:

- Domains of interest (e.g. politics, religion, etc.) that are mined by learning dedicated models from external resources such as Wikipedia.
- Location trail, including home location and visited places that are estimated by using probabilistic location models from large quantities of geolocated training data.
- Favorite brands and products (e.g. mobile phones, clothes) that are mined through the identification of named entities (brand and product names) which appear in users' texts.
- User's stance on their areas of interest that is extracted using opinion mining tools which are adapted to a use with OSN multilingual and heterogeneous content.
- Social affinities (i.e. people sharing similar content) that are discovered via a comparison of user's contributions in different domains of interest.

A variety of personal information is shared on OSNs, including: (a) status updates added by the users, (b) comments on their multimedia content (i.e. photos, videos) and (c) third-party publicly available texts (i.e. newspaper articles, blogs etc.) that are shared by the users. Due to this variety of information, flexible and extensible text mining and linking modules and approaches need to be employed. During the first iteration of the project, we prioritized the implementation of multilingual text similarity and location detection functionalities due to their importance for the privacy framework dimensions developed as part of WP6. The first stream

of work focused on adapting a **text similarity** (Section 2) method, which can be used to compare texts across languages. Support for multilingual text processing is important for two main reasons: (1) the structured resources needed for text mining are often difficult to build in a large number of languages due to the scarcity of raw data in these languages, and (2) a user may share content in different languages. A second important approach is **location detection** (Section 3), which attempts to estimate the location(s) associated with a piece of text shared on OSNs.

1.2. Research methodology and contributions

Research on text mining and linking is part of the multidisciplinary research effort required by USEMP use cases and it is thus largely shaped by the conclusions of upstream research from other disciplines (notably legal studies, user studies and system design). It is also closely interlinked with visual content mining since text and image modalities often offer valuable complementary insights, which can be jointly exploited to improve the mining process. In USEMP more focus is put on visual content mining, which is more challenging, and a choice was made in the project's DoW to adapt a majority of text mining modules from existing approaches that are aligned with the USEMP objectives and are also well mastered by project partners. After a careful analysis of Intellectual Property (IP) rights, existing implementations of text mining were reused wherever possible. To assess the reliability and quality of the prototyped solutions, they were evaluated using suitable publicly available datasets and in the case of location estimation, through participation in an international benchmarking activity.

Although much of the work performed in this first research and development iteration relied on existing text mining approaches, we consider that it resulted in valuable research contributions concerning location recognition and text similarity. More precisely, the experiments carried in this area showed that a careful combination of probabilistic models learned with large scale training data and of social cues significantly improves results compared to state of the art techniques. Concerning text similarity, the main contribution is related to the adaptation of the method to domains that are relevant in privacy contexts.

1.3. Multidisciplinary issues[1]

Although text mining is mainly dealing with approaches from natural language processing and machine learning, the presented research was considerably shaped by the rest of the USEMP disciplines, and at the same time provides actionable feedback to them. In the following, we provide a concise account of the inter-play between text mining research and the different disciplines of the project.

D5.2 is informed by work done in WP2, WP3, WP4 and WP9 and it provides valuable input for WP6 and WP7. The legal analysis carried out in WP3, and more particularly in T3.6 which deals with the coordination of legal aspects, clarified practical implications of text mining related to: processing of sensitive information, copyright issues related to data used during training, and ensuring that all USEMP components have clear IP rights (in case of reusing existing components). Work on trade secrets and intellectual property done as part of D3.2

[1] Multidisciplinary issues are, to a large extent, common to all WP5 deliverables and this section has thus similar content in D5.1, D5.2 and D5.3.

explored the tensions between profile representations on the end-user side, within OSNs and created in USEMP and made clear the complex interplay between these actors, as well as their respective rights and obligations.

The use case analysis in D2.1 and the associated requirements defined in D2.2 served as guidelines for the implementation of technical components. In particular, the following requirements are central here:

- [SR02] “The system may be able to process the information within one second such that the user can make informed decisions on their past data without long delays. In the event data processing is to take longer, a progress bar should be presented. A maximal extent of 10 seconds will be aimed for.” This requirement has strong implications in terms of processing speed for the implemented components.
- [SR04] “The system may be able to make best effort associations between data placed onto OSN(s) and the profile attributes which can be inferred from such data.” This requirement is a counterpart of [SR02] that focuses on component performance, which should closely follow state of the art developments.
- [SR11] “The system may be able to get fruitful insights on how relevant a user’s profile is for different stakeholders.” Through inferences made by technical components, the end-users should be able to have insightful information on how her profile is seen by OSNs and, possibly, by other stakeholders.

In D4.1, a comprehensive list of social requirements was established, which offers a user-side view of the expected behavior of the developed USEMP tools. Of particular interest here are:

- Req. 1 asking for more transparency about privacy problems at an institutional level and notably OSNs in this context.
- Req. 2 demanding a backward link between inferences and raw data which generated them to improve the explainability of the automatic decisions made by the system.
- Req. 10 asking for a low impact on browser speed of the USEMP plug-in, a requirement which is tightly linked to [SR02] mentioned above.

The extensive market analysis done in D9.3 showed that existing privacy enhancing tools and privacy feedback and awareness tools deal mostly with volunteered and/or observed data. A strong opportunity in USEMP is to provide users with a more complete view of how their data could be handled and exploited by OSNs. Another conclusion of D9.3 is that existing text mining tools are not tailored for privacy enhancement and, consequently, an adaptation step is needed in order to better satisfy domain requirements. Downstream, insights gained with D5.1 tools can be used both directly in the USEMP interface (D7.2), and as part of the privacy scoring framework created in D6.1, to complement social network mining inferences. For instance, user locations can be extracted from texts and can be displayed directly by the USEMP interface to inform the user about her degree of exposure on a certain privacy dimension (e.g. political beliefs). In a more complex functioning mode, text representations can be compared to the Facebook pages gathered and structured as part of WP6 in order to derive domains of interest for the users. Furthermore, they can be combined with social interaction data (such as likes, comments) to improve the quality of predictions.

2. Text similarity

2.1. Related work

Explicit Semantic Analysis (Gabrilovich & Markovitch, 2007) is a method that maps textual documents onto a structured semantic space. Since its introduction in 2007, ESA was successfully exploited in different NLP and IR tasks. The success of this simple method lies in the richness and the quality of the underlying conceptual space but also in its ability to provide results which are easily explainable for the end-user. In the original evaluation, ESA outperformed state of the art methods in a word relatedness estimation task and different developments were subsequently proposed. (Radinsky et al., 2011) added a temporal dimension to ESA vectors and showed that this addition improves the results for word relatedness. (Hassan & Mihalcea, 2011) introduced Salient Semantic Analysis, a variant of ESA that relies on the detection of salient concepts prior to linking words and concepts. The merits of their method are difficult to estimate since the comparison is often made with an in-house ESA implementation whose results are significantly poorer than those presented in (Gabrilovich & Markovitch, 2007). (Popescu & Grefenstette, 2011) proposed an ESA adaptation to IR tasks that gives priority to categorical information. The comparison with a classical ESA implementation showed that a significant improvement was obtained in an image retrieval setting. Moreover, the method compared favorably with other state of the art indexing and retrieval schemes in an ad-hoc image retrieval task. ESA has only weak language dependence and was already deployed in several languages. (Sorg & Cimiano, 2012) proposed an extension of the method to different languages and showed that the method is useful in cross-lingual and multilingual retrieval settings. (Bouamor et al., 2013) introduced an approach that adapts ESA representation to different domains by domain seeding, which necessitates minimal manual intervention. This method allows one to select only a subset of the ESA conceptual space, which can then be used to compute the similarity between a test document and a domain of interest.

Recently, deep learning methods achieved very promising performance in different tasks and they were also applied to derive word similarities (Huang et al., 2012). Their performance in word similarity estimation is slightly lower than those of ESA-inspired methods but text representations are more compact. This last property makes them interesting from a scalability perspective but their adaptation for multilingual contexts, such as the one in USEMP is not straightforward and goes beyond the scope of the project.

2.2. Method description

Here, we propose to modify the classical formulation of ESA in order to better cope with short texts. We notably create ESA models for the four languages of interest in USEMP, namely English, Dutch, Swedish and French.

2.2.1. Classical formulation of ESA

ESA exploits classical text weighting schemes, such as the tf-idf, to model concepts from a structured resource, such as Wikipedia where each article is an unambiguous concept. The weighted representation of a Wikipedia concept is expressed as:

$$C_x = ((T_1, \quad w_1^x), (T_2, \quad w_2^x), \ldots, (T_N, \quad w_N^x))$$

with *N* - the size of the term vocabulary used to model Wikipedia content, T_k - the k^{th} term from this vocabulary and w_k^x - the weight of term T_k for concept C_x. Knowing that the size of term vocabularies is in the range of 10^5 and that Wikipedia articles usually include a few hundreds of distinct terms (Gabrilovich & Markovitch, 2007), C_x representations will be sparse (i.e. most w_k^x will be 0). Consequently, the relation between the words and the concepts that span the space can be efficiently written using an inverted index structure that stores only non-null w_k^x values. Thus, each word T_k of the vocabulary has an associated high-dimension projection onto the concept space of the underlying resource, which can be written as:

$$T_k = ((C_x, \quad w_k^x), (C_y, \quad w_k^y), \ldots, (C_z, \quad w_k^z))$$

where the number of active (non-zero) concepts C_x is much smaller than the total size of the conceptual space defined by Wikipedia.

In order to obtain the similarity between two documents, their ESA representations are obtained by summing up the weights of individual concepts which are associated to document terms T_k and the aggregated concept weight of C_x for document D_1 can be expressed as:

$$w_{D_1}^x = \sum_{k=1}^{P} w_k^x$$

with *P*, the total number of terms in document D_1.

Finally, the similarity between two documents D_1 and D_2 is computed as the dot product between the weights of the concepts C_x which are associated to both documents:

$$sim(D_1, D_2) = \sum_x w_{D_1}^x * w_{D_2}^x$$

Classical ESA representations are well adapted for single words (limit case where the length of both documents $P = 1$), since comparison of ESA vectors can be done directly, and for long documents, since the summing operation leverages individual contributions and an accurate semantic representation of the document is obtained. However, the method has some drawbacks for documents such as short comments and messages posted to OSNs (with document length $P \leq 10$). Here, the smoothing of individual contributions is not sufficient because the contribution of a single word can be higher than that of the others and the obtained related concepts could be related to a part of the topic only. An illustration of this type of problem is provided in Table 1, which presents the top 10 ESA concepts associated to “freshwater fish”, “Jean-Jacques Rousseau” and “crockery doll house”. The results from Table 1 indicate that most ESA top ranked concepts are not related to the entire query. When examining results for “freshwater fish”, “Freshwater bivalve” and “Freshwater, Isle of Wight” are related to “freshwater”, while “Bait fish” and “Bank fishing” are related to “fish”. Similarly, when examining results for topic “Jean-Jacques Rousseau”, we notice that several ESA top concepts are brought up by the family name “Rousseau” and have reduced semantic relatedness with the original topic. Concepts found for “crockery doll house” are related to doll but not the other terms from the query.

Input text	Top 10 classical ESA concepts	Top 10 classical adapted ESA concepts
freshwater fish	Freshwater bivalve; Freshwater mollusc; Tropical fish; Freshwater, Humboldt County, California; Fish fillet processor; Bait fish; Fish marketing; Bottom fishing; Freshwater, Isle of Wight; Bank fishing	Eastern freshwater cod; Ide (fish); New Zealand longfin eel; Common galaxias; European perch; Green swordtail; Rainbowfish; Common rudd; Spotted bass; Common bream
Jean-Jacques Rousseau	Confessions (Rousseau); Saint-Jean; Considerations on the Government of Poland; Eugène Rousseau (chess player); John Jacques, Baron Jacques; Eugene Rousseau (saxophonist); Jean-Jacques Henner; Victor Rousseau; Bobby Rousseau; Discourse on the Arts and Sciences	Confessions (Rousseau); Considerations on the Government of Poland; Discourse on the Arts and Sciences; Emile, or On Education; Essay on the Origin of Languages; Discourse on Inequality; Letter to M. D'Alembert on Spectacles; Pygmalion (Rousseau); Julie, or the New Heloise; Le devin du village
Crockery doll house	Peg wooden doll; Composition doll; Anatomically correct doll; Bisque doll; Black doll; Paper doll; Madame Alexander; Fashion doll; Doll ; China doll	Mabel Lucie Attwell; Bringing Up Father; The Tale of Mrs. Tiggy-Winkle; China doll; Japanese traditional dolls; Queen Mary's Dolls' House; Bild Lilli doll; Vivien Greene; Paper Dolls (band); Wall House (Elkins Park, Pennsylvania)

Table 1. Top 10 ESA related concepts for "Freshwater Fish", "Jean-Jacques Rousseau" and "crockery doll house". The second column contains results for classical ESA (Gabrilovich and Markovitch, 2007), while the third column present results for the adapted version of ESA introduced in (Popescu, 2013).

2.2.2. ESA adaptation for short texts

In (Popescu & Grefenstette, 2011), we proposed a version of ESA that gives a privileged role to categorical information. The method included two scores to rank Wikipedia concepts appearing in ESA text representations:

- a Boolean score that captures the number of common words between the initial topic and the words found in the categories associated to Wikipedia concepts.
- the score used in the classical ESA in order to rank concepts, based on the sum of the contributions of the individual words.

Since OSN texts are often short, ties are often obtained with the Boolean score and they are broken using the second, finer-grained score.

The introduction of the Boolean score has two main objectives. First, categorical information should be favoured in order to obtain concepts that are hierarchically related (i.e. *Is-A* relation) to the initial topic or to parts of it. Second, it is possible to identify which parts of the initial short text an ESA related concept is related to. For instance, the categories of concept "tropical fish" from Table 1 are "fish stubs" and "aquaria" and this concept would have a Boolean score of 1 out of a maximum of 2 for the input text "freshwater fish". Similarly, "freshwater bivalve", the top ranked concept with classical ESA and only loosely related to the initial topic, has a Boolean score of 0 since its only category is "bivalves". The categorical ranking rightly gives a better position to "tropical fish" compared to "freshwater bivalve" since the first concepts is more closely related to "freshwater fish".

In (Popescu, 2013) we further modified our ESA adaptation for IR in two directions. First, given that categorical information is often sparse, we added the words contained in the first

150 characters after the concepts name in the first paragraph of the category words. This enrichment of the categorical space is motivated by the fact that the first paragraph of a Wikipedia article is often a definition that contains salient concepts related to the target one. The limitation to the words contained in a string of 150 characters is useful since the first paragraph has varying length and contains information that is only loosely related to the concepts when it is long. Assuming that an ESA index was created from the categories and the first sentence, a similarity score is computed based only on this index and is expressed as:

$$sim_c(D_1, D_2) = \sum_x cw_{D_1}^x * cw_{D_2}^x$$

Where $cw_{D_1}^x$ and $cw_{D_2}^x$ are term weights obtained only from categories and first sentences. $sim_c(D_1, D_2)$ can be used to alter the initial similarity score $sim(D_1, D_2)$ between documents D_1 and D_2. For instance, the two similarity values can be multiplied to boost the similarity score of documents that share categorical information.

The second modification is a concept detection that is used to produce a third score that favours articles that contain longer concepts from the initial query over other articles. At equal categorical scores, the inclusion of concept detection allows us to favour a Wikipedia concept that includes “Jean-Jacques Rousseau” in its text when compared to another concept that includes “Jean-Jacques” and “Rousseau” separately. The top related concepts obtained with adapted ESA for “freshwater fish” and “Jean-Jacques Rousseau” are presented in the third column of Table 1. In both cases, the top 10 concepts are much more closely related to the initial topics compared to the use of classical ESA. The list of related concepts for “freshwater fish” contains only fish and the list for “Jean-Jacques Rousseau” includes different works of the philosopher. “Crockery doll house” is a specialized text, which is not well represented in Wikipedia and the retrieved concepts are still unrelated to the entire topic in a large majority of cases. This last topic illustrates one limitation of any ESA implementation, namely the poor mapping between the initial document and the knowledge included in the underlying conceptual space.

2.2.3. Multilingual ESA representation

Inspired by existing work in (Sorg & Cimiano, 2012) and (Popescu & Grefenstette, 2011), we exploit the Wikipedia translation graph in order to map texts in a common representation space. Given the richness of resources available in English, we use this language as pivot and translate the ESA text representations from any language into English. Although incomplete, the Wikipedia translation graph enables a relatively clean translation of unambiguous concepts. Furthermore, this translation facilitates comparability of texts by expressing all of them in English. For instance assuming that a text in French mentions “Jean-Jacques Rousseau”, all associated French Wikipedia concepts with English translations will be used to represent the initial text in English.

2.2.4. Domain adaptation of ESA

As we mentioned, ESA modelling was chosen because they cover a large spectrum of domains and can thus be used to represent highly diversified content shared on OSNs. In USEMP, we are mainly interested in domains that match privacy dimensions defined as part of D6.1. Inspired by prior work in (Bouamor et al., 2013), we propose a domain adaptation of ESA in which only concepts that are closely linked to a given privacy dimension are retained

in the representation. Relevant domain concepts are selected by comparing them to a domain prototype which is extracted from a prototypical document which is manually defined. For instance, if we want to model "politics" in English and Dutch, the prototypical documents can be http://en.wikipedia.org/wiki/Politics and http://nl.wikipedia.org/wiki/Politiek. Weighted representations of these documents that include only the highest weighted terms are expressed as:

$$proto_{dom} = ((T_1, \quad w_1^{dom}), \quad (T_2, \quad w_2^{dom}), \ldots, (T_{||dom||}, \quad w_{||dom||}^{dom}))$$

with $||dom||$ the number of distinct terms used to represent the domain, typically in the range of 10 - 20.

Inspired by (Bouamor et al., 2013), $proto_{dom}$ is used as domain prototype and the other Wikipedia documents in each language are sorted by computing their domain ranking of concepts C_x as follows:

$$rank_{dom}(C_x) = \left(\sum_{k=1}^{||dom||} w_k^{dom} * w_k^x \right) * count(proto_{dom}, C_x)$$

with w_k^{dom} – the weight of term T_k in the prototypical domain representation $proto_{dom}$, w_k^x – the weight of term T_k in the concept to rank C_x, and $count(proto_{dom}, C_x)$ – the number of distinct terms from the prototypical domain representation which also appear in the concept representation C_x. The first term of the equation sums up the contributions of different words from the candidate concept C_x, while the second is meant to further reinforce articles that contain a larger number of terms from the domain representation.

$rank_{dom}(C_x)$ expresses the similarity between a concept and a domain of interest and only the top ranked concepts are retained as representative. The weights of articles from the initial ESA representation whose domain scores are below the threshold are forced to zero when computing the similarity between a text and a privacy domain.

With $rank_{dom}(C_x)$ available, we can create vectorial domain representations by exploiting the tf-idf modelling of articles that are most relevant for each domain (typically up to 10,000 articles per domain). A list of vocabulary of terms with highest document frequency (i.e. number of distinct articles in which they appear) in Wikipedia is created and used as a representation space for domains. Typically, this vocabulary includes 10,000 to 20,000 terms in order to keep a good balance between vocabulary expressivity and computational complexity of text similarity computation. Domain models are obtained by summing-up the tf-idf scores of vocabulary words in the domain-related articles. The similarity between an input text and domain models can then be computed by using similarity measures such as the dot product or the cosine similarity.

2.3. Evaluation and testing

2.3.1. Preliminary experiment

To validate our implementation, we first performed the word similarity task described in (Gabrilovich & Markovitch, 2007), with the same version of Wikipedia, and the method achieved a 0.72 correlation with human judgments (to be compared with 0.75 reported by (Gabrilovich & Markovitch, 2007). This difference is probably due to minor implementation

differences but, as we mentioned, also to the fact that some implementation details were not available until recently.

2.3.2. Semantic enrichment

A second evaluation, which was part of the CLEF CHiC 2013 evaluation campaign[2], was realized in a semantic enrichment context whose aim is to assess the accuracy of methods which provide a list of related concepts for an input text. In the case of ESA, this evaluation allows us to compare the efficiency of classical and the adapted implementation of the method:

- adapted ESA[3] - Related concepts are obtained with the adapted version of ESA.
- classical ESA - Related concepts are obtained with classical ESA.

The evaluation was run over 25 diversified short English texts that were manually judged by a pool of experts. For each text, the top 10 associated concepts produced by automatic methods were retained for relevance judgment, and labelled as being “irrelevant”, “partly relevant” or “relevant”. The results obtained with the two ESA variants and with a link-based method proposed by (Lam Tan et al., 2013) are presented in *Table 2*. These results show that both ESA-based methods outperform a Wikipedia link-based method. More importantly, the adapted version of ESA improves results by over 100% for a strict measurement of P@10 and by 90% for a loose measurement when compared to the classical formulation of the method.

Method	P@10 strict	P@10 loose
adapted ESA	0.468	0.66
classical ESA	0.212	0.364
(Lam Tan et al., 2013)	0.16	-

Table 2. Semantic enrichment accuracy measured using the Precision at 10 (P@10) measure. Strict refers to the case when only “relevant” judgments are considered for precision calculation and “loose” refers to the case when “partly relevant” results are also considered. (Lam Tan et al., 2013) propose a method which is also based on Wikipedia but is focused on the exploitation of article links. Higher scores are better.

2.4. Implementation and usage

We use an in-house implementation of ESA that includes only the optimization cues publicly available until recently[4]. Our implementation of the ESA models computation was made in Perl, with a focus on its adaptability to a large array of languages. A configuration file allows users to add new languages whenever needed. Given their sparsity, the ESA indexes are written as an inverted index that links words to an array of concepts (and their associated weights). Separate indexes are written in each language for the categorical information and for the entire articles. The creation of ESA indexes is done offline and parallelization on CPUs is used to speed-up the process.

[2] http://www.promise-noe.eu/chic-2013/home (consulted on 31/12/2014)
[3] Here “adapted” refers to adaptation to short texts and not to domain adaptation.
[4] A full list was recently made public at https://github.com/faraday/wikiprep-esa/wiki/roadmap but the remaining cues were not yet integrated in our implementation.

In USEMP we expose the module that queries ESA indexes to obtain text representations and then translate them in English. This module is also written in Perl and can be called with the following command:

The text similarity computation wrapper can be called with the following command:

> compute_similarity [num-domains] [num-dimensions] [domain-models] [tmp-ascii-l2] [tmp-domains] [top-domains]

The commands and parameter files are explained in Table 3. Text similarity program description and usage.. This extraction assumes that the Caffe suite is already running on the server, with GPU enabled and that the same CNN model used to create concept models is readily available.

Program	**Description**
compute_similarity	C++ binary used to compute the most salient concepts of an image.
Arguments	**Description**
num-domains	Number of domains modeled in USEMP and for which we can compute a similarity with the input text.
num-dimensions	Number of dimensions of the domain models.
domain-models	Precomputed domain models written in a flat text format. Dimensions are separated by simple space.
tmp-ascii-l2	Input text in liblinear format.
tmp-domains	Output file which stores the list of most relevant concepts for the current image. Concepts are ranked by decreasing classification score.
top-domains	Number of most salient concept retained for each image.

Table 3. Text similarity program description and usage.

The implementation is mature and minor effort is needed in order to integrate concept detection in the USEMP system. Consequently, this brick will be used from the very beginning of user tests which are due to start on February 2015.

2.5. Next steps

The experiments validate our implementation of ESA and also show that the adapted version that privileges categorical information has better behavior in a semantic enrichment task. Future work will be carried out in close collaboration with WP6 and will mainly involve the refinement of domain adaptation in order to propose finer grained insights into a user's degree of exposure in a given domain. The user will thus be able to understand not only how exposed she is on a particular dimension but also which concepts trigger that exposure. After the association of a document to a domain, concepts from the domain that are explicitly present in the document will be extracted and presented to the user. The document-domain association is important in order to reduce ambiguity whenever a concept is ambiguous in Wikipedia. This concept detection approach is interesting because it includes the detection of named entities and of other concepts, but also because it entails the linking of text concepts to unambiguous Wikipedia concepts. Another important challenge refers to the fact that the Wikipedia translation graph is incomplete and valuable information is lost during translation. To cope with this problem, we will investigate the use of concept similarity matrices in each language in order to represent each ESA concept as a distribution of related concepts and thus improve comparability across languages.

3.Location detection from texts

In USEMP, location detection from texts is important in particular for building the user location profile, which is one of the eight core privacy dimensions determined through work done in WP4 and WP6. Given that location prediction from texts is more reliable than the one based on images, this tool will provide the main inputs for building the user's location profile. However, its fusion with location detection from images (see D5.2), which is likely to improve the overall quality of predictions, will be explored in the dedicated fusion deliverable (D5.3). Location detection is thus mainly useful for the first use case of the project as it contributes to raising the user's awareness about what can be automatically extracted from her data. In USEMP, we have mainly investigated the use of robust probabilistic models that constitute an improvement over existing approaches described in (Serdyukov et al., 2009) and (Popescu & Ballas, 2012), with contributions related to improving the robustness of the models. A second important finding concerns the positive role of social cues (i.e. special tags, user models) as a complement to the plain use of probabilistic models.

3.1. Related work

Linking texts to locations was traditionally regarded as a named entity recognition task done with the help of extensive geographic databases but, since the seminal work of (Serdyukov et al., 2009), more focus was put on the use of probabilistic language models adapted to the geographic domain. The creation of such models was enabled by the availability of large quantities of geolocated metadata on Web 2.0 platforms such as Flickr, Twitter and Wikipedia. Put simply, the world surface is split in (nearly) rectangular cells in order to compute the intensity of the link between terms and cells by exploiting language modelling approaches from information retrieval (Ponte & Croft, 1998). In follow-up work, (O'Hare & Murdock, 2013) investigated the role of social cues in improving the quality of probabilistic location models. They notably found that the use of user counts instead of raw photo counts constitutes an efficient way to reduce the negative effect of bulk uploads and, consequently, to increase the robustness of the obtained models. Equally interesting, they show that using a matching of a text with a given cell of the probabilistic model and with a low-weighted representation of neighboring cells improves overall results. (Van Laere, et al., 2012) explored the use of clustering models to find a more semantically grounded division of the world surface. While interesting, this method has the disadvantage of poor scalability due to the complexity of the clustering step. More recently, (Van Laere et al., 2014) showed that location models learned on Twitter and Flickr can be transferred for predicting the location of Wikipedia articles, although the nature of these documents is different. However, they also show that Flickr is more reliable compared to Twitter, probably due to the fact that, in Twitter, many documents are geolocated by default but do not actually have any geographical intent. Consequently, such documents will act as noise when included in the training sets used to derive the location models.

3.2. Method description

The text-based location detection method implemented in USEMP is mainly based on the use of probabilistic location models. These models are complemented with the use of social cues (i.e. location related machine tags and user geolocation models) whenever the latter

are likely to provide a more reliable prediction. We describe each component separately, as well as a cascade fusion that was implemented to combine them.

3.2.1. Probabilistic location models

The problem to solve here can be expressed as: given an arbitrary text, provide its most probable location (i.e. pair of geographic coordinates or place name) based on a modelling of the physical world, which is split in cells (i.e. nearly rectangular region of the world) whose size is approximately 1 km^2. This size of cells is empirically motivated by the fact that, in a large majority of cases, it is difficult to predict location with a smaller error (Popescu & Ballas, 2012). We assume that a large amount of geolocated textual metadata are at hand and can be used for training. As we mentioned, in USEMP we tackle this problem by using location models inspired by (Serdyukov et al., 2009), with two notable differences:

- Term counts are replaced by user counts in order to counter the negative effect of bulk uploads. This is important since counts can be skewed by a large number of contributions from a single user in a cell. For instance, if a user uploads 1000 photos wrongly annotated with “Versailles” somewhere in China and term counts are used, the importance of Versailles in this cell will be overweighted. If user counts are used instead, the overweight will be removed since all photos are assigned to a single user.
- Language models used in (Serdyukov et al., 2009) are replaced by a simpler probabilistic representation of terms that is more scalable and also does not require a validation phase to tune parameters. The probability of a term in a cell is simply obtained by dividing its user count in that cell to its total user count in all cells of the model.

Probabilistic language models can be built from any geolocated resources. Typical data sources include Twitter and Flickr (Van Laere, 2014), with short texts associated to pairs of geographic coordinates. The advantage of such sources is that, given the limited amount of information, it is likely for them to be focused on particular locations while in longer texts, such as Wikipedia articles, parts of those texts are likely to be unrelated to the actual location being modelled. Given its availability and its better performance compared to Twitter (Van Laere, 2014), we chose to create models based on Flickr metadata. The model of a cell can be expressed as:

$$C_x = ((T_1, \quad w_1^x), (T_2, \quad w_2^x), \dots, (T_N, \quad w_N^x))$$

with *N* - the size of the term vocabulary used to model locations, T_k - the k^{th} term from this vocabulary and w_k^x - the probability of term T_k in cell C_x.

Since cell representations are sparse, with only a minority of terms having non-null probabilities in each cell, an inverted index representation of the models is used in order to accelerate the prediction phase. Consequently, the models are rewritten as:

$$T_k = ((C_1, \quad w_k^1), (C_2, \quad w_k^2), \dots, (C_M, \quad w_k^M))$$

with *M* the total number of cells modelled.

Given a test text which includes an arbitrary number of terms from the modelled vocabulary, the score of a cell is computed by summing up each term’s probability in that cell. Then cells are sorted in order to find the most probable one, which will be the assigned location of the tested text.

3.2.2. Location related machine tags

Machine tags are triples that are composed of a namespace, a predicate and a value and allow a user to unambiguously express knowledge about an event, a location, an object, etc. In our case, we are mainly interested in machine tags that are strongly associated to locations. Applying this constraint, after an initial analysis of more than twenty different types of machine tags, we selected only Foursquare, Last.fm and Upcoming entries and integrated them in our location prediction framework. For each such tag, we use the training set to learn its most probable location and then exploit it during the test phase.

3.2.3. Geolocation user models

If images do not have associated tags or if these tags are not geographically discriminant, placing photos with probabilistic models is likely to fail. To overcome this problem, we exploited a simple user modeling technique, which computes the most probable cell of a user if a number of geolocated items are already associated to that user.

3.2.4. Geolocation procedure

The geolocation methods presented in the preceding sections are used in a cascade combination. Initial experiments showed that machine tags are very reliable and are used first if they exist. However, they are seldom present and a choice between probabilistic models and user models is needed. To make this choice, we empirically determine a threshold using a validation set. If the score of the most probable cell is above the threshold, probabilistic models are used. Otherwise, user models are exploited.

3.3. Evaluation and testing

We tested our approach in the MediaEval Placing Task 2014 challenge[5], where the main objective was to estimate geographical coordinates of multimedia items, such as images and videos, based on massive amounts of geo-tagged training data. The challenge was run using Flickr data extracted from the recently released YFCC dataset[6], with over 5 million items given for training and 510,000 items for testing (all specified by the organizers). In addition to this dataset, we exploited external data to verify the assumption that better results are obtained with the use of more data. More precisely, we exploited: (1) all geotagged metadata from the YFCC dataset after removing all test items and (2) an additional set of ~90 million geotagged metadata from Flickr.

As we mentioned, the surface of the earth was split in (nearly) rectangular cells of size 0.01 of latitude and longitude degree (approximately 1km^2 size). Both titles and tags were taken into account and are referred to as tags hereafter. For user models, we downloaded up to 500 geotagged images per user in order to determine her most probable cell. In addition, only photos that are at least 24 hours away from any of the user's test set images were exploited to reduce the risk of learning from test data.

We submitted the following runs:

- RUN_1 - exploited location models and machine tags from internal training;

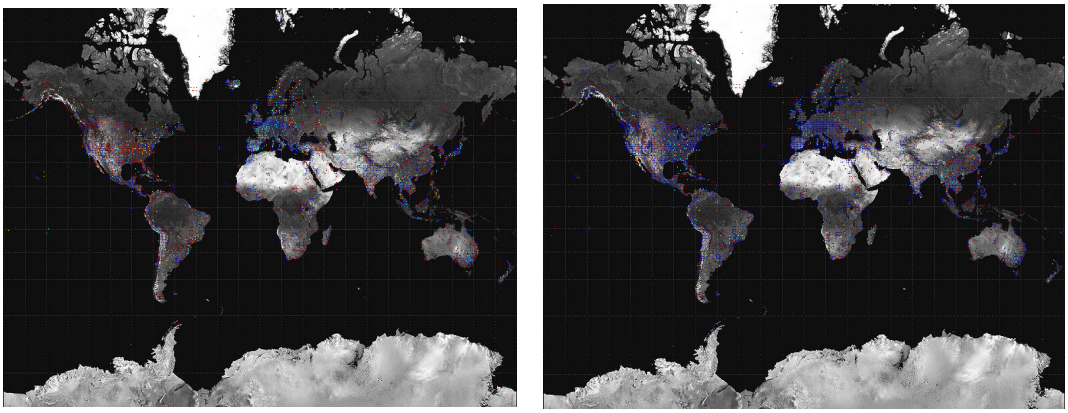
[5] http://www.multimediaeval.org/mediaeval2014/placing2014/ (consulted on 29/12/2014)
[6] http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67 (consulted on 29/12/2014)

- RUN_2 - combined location models and machine tags from the entire geotagged YFCC dataset, after excluding test items;
- RUN_3 – exploited location models from the entire YFCC dataset and user models;
- RUN_4 - used YFCC and fusion of location models, machine tags and user models.

We present the performance of the submitted runs in Table 4. As expected, the best results were obtained when combining all types of available information (RUN_4). The analysis of the different runs also shows that the largest contribution is due to probabilistic location models. The large gap between RUN_1 and the others (close to 100% improvement) confirms that the use of supplementary training data is very beneficial. This difference is illustrated in Figure 1, with the precision results obtained with the official training set of Placing Task (a) and a larger training set (b). As expected results are better in regions that receive a lot of user contributions, such as Western Europe or the East and West coasts of the United States. Interestingly, geolocation of items from the United States is overall quite challenging and this is probably due to the fact that many place names are highly ambiguous in this region of the world. For instance, the Geonames[7] database includes 30 different cities, towns and villages named “Geneva” in the United States, while the best known city with this name is in Switzerland. Without disambiguation any photo tagged with Geneva in one of the US places would automatically be placed in Switzerland.

Run	P@0.1 km	P@1 km	P@10 km	P@100 km	P@1000 km
RUN_1	0.016	0.235	0.408	0.481	0.618
RUN_2	0.043	0.428	0.582	0.644	0.753
RUN_3	0.012	0.418	0.597	0.679	0.779
RUN_4	0.043	0.441	0.613	0.691	0.786
Other best	0.044	0.222	0.390	0.461	0.599

Table 4. Geolocation prediction performance on the MediaEval Placing Task 2014 dataset. Accuracy is measured using precision at different granularity levels, according to the requirements of the task. P@X km number of test items placed at least than X kilometers from their true location. Naturally, higher scores are better. “Other best” stands for the best result reported by (Kordopatis-Zilos et al, 2014), which was also based on a probabilistic language model.



(a) (b)

Figure 1. Automatic geolocation precision for the MediaEval Placing Task 2014 with different sizes of the training set: 5 million metadata pieces in (a) and 45 million in (b). Red and blue dots correspond to bad and good geolocation precision respectively.

[7] http://www.geonames.org/ (consulted on 30/12/2014)

The difference of precision at close range (P@0.1) between RUN_2 and RUN_3 confirms that machine tags are very useful for precise geolocation. While their usefulness for test datasets outside of Flickr is limited, this finding indicates that whenever special terms appear in a text, they should be exploited. In the case of Facebook, these special terms can be links to location-related pages or hashtags. User models are only useful if larger errors (above 1 km) are admitted and this result was expected since they only give a rough indication about the most probable location of a user. In Facebook, user models could be created from the existing user contributions or be directly available through her home location if this information is provided. We will produce such models as soon as Facebook data become available during USEMP user tests. The combination of all available approaches in RUN_4 gives the best performance for all precision ranges.

In addition to the official runs, we have also explored the creation of models from an even larger dataset and added another ~90 million metadata pieces that were had been collected by the Georama project[8] before 2010. Surprisingly, the use of these supplementary data had a slightly negative effect on results (approximately 1 P@1km point lost when compared to the best configuration in RUN_4).

3.4. Implementation and usage

A Java implementation of the probabilistic location modelling approach of section 3.2.1 has been delivered. No use of machine tags and user modelling is made since machine tags are not used in Facebook posts and since no training data is available from Facebook to assess the effect of a user modelling approach. Instead, the implementation includes an extension described in (Kordopatis-Zilos et al., 2014), which makes use of a dual cell grid with coarser and finer structure in order to enable higher location estimation accuracy in fine ranges (below 1km).

Together with the implementation, we make available a pre-computed location model. Since the model was generated for many millions of Flickr image metadata, several filtering operations were applied in order to make the model more compact. Still, the model requires approximately 4GB of main memory to be fully memory-based. This amount of memory is currently considered standard; hence no effort was spent on optimizing for memory usage.

To generate location predictions for a set of input text messages, the following command should be executed (assuming a JRE is installed):

```
java -Xms4G -Xmx4G -jar geopred.jar [root-folder] [text-input-path] [output-path]
```

where [root-folder] denotes the folder where the library and location model files reside, [text-input-path] is the path to a text file containing the input texts (one per line), and [output-path] is the path to a text file containing the predicted locations (one per line), which have the form:

latitude longitude city country (tab-separated)

At the moment, executing the above command loads the location model in memory and then performs the prediction. For sets of input texts that are small or moderate in size (e.g. a few hundreds to thousands), this is clearly an unacceptable overhead (since loading the location model in memory typically takes between two and three minutes). Hence, prediction should

[8] http://www.kalisteo.eu/en/project.htm

either be performed in large batches of texts (tens to hundreds of thousands) or the module should be exposed as a service (i.e. the location model will be pre-loaded in memory and requests will be served using the pre-loaded model).

This implementation is now stable and will be integrated in the USEMP system for the pre-pilot tests to be carried out in February 2015.

3.5. Next Steps

The experiments carried out with the Placing Task 2014 dataset give very interesting results and, given the high importance of location among privacy dimensions (see user study in D4.3), work will be carried out to further improve this module. We will notably explore ways to combine textual and visual geolocation as part of D5.3 but also focus on the intrinsic quality of probabilistic location models. One direction which was not explored in literature and seems promising is to evaluate the geographic relevance of textual annotations of geotagged items before including them in the training set. From an integration point of view, we will provide the text-based location detection library for inclusion in the USEMP system.

4.Conclusions and future work

During the first iteration of the project, work on developing textual mining and linking modules was conducted in two main directions – text similarity and location detection from texts. Competitive text representations were devised in both cases by leveraging recent advances in text mining. For text similarity, we implemented a version of Explicit Semantic Analysis (ESA), a method that addresses well the USEMP requirements in terms of wide conceptual coverage and multi-linguality. We have shown that a version of ESA that gives a privileged role to categorical information clearly outperforms the classical formulation of the method in a semantic enrichment of short texts task. Equally important, we have introduced a domain-adaptation of ESA which allows the text mining module to make a link between users' texts and privacy dimensions defined as part of WP6. A challenge that will be tackled in future work is the incompleteness of translated ESA text representations. For location recognition, we introduced a simple and scalable formulation of probabilistic location models and combined them with other cues. Experimental validation done as part of the MediaEval Placing Task 2014 showed that our method clearly outperforms the systems submitted by other participants. Future work on location detection will focus on the quality of raw data that are included in the training set. Equally important, the fusion of text and image modalities for improved location detection will be studied as part of D5.3.

Two important modules will be added to the processing framework during the second iteration: a) a concept detection method that includes both named entities and other concepts and exploits the domain adaptation capabilities of ESA for concept disambiguation and linking, and b) a sentiment analysis method that enables text mining beyond a neutral characterization.

In parallel to improving and extending the text mining modules, we will focus on their integration in the USEMP system. As we mentioned, we will first integrate the location detection module for the pre-pilot studies to be carried out in February 2015. Then, we will progressively integrate the other modules with the overall objective of reaching full integration by the end of the second reporting period (September 2015).

5.References

D. Bouamor, A. Popescu, N. Semmar, P. Zweigenbaum (2013) Building Specialized Bilingual Lexicons Using Large-Scale Background Knowledge. Proc. of *EMNLP 2013,* Seattle, USA.

E. Gabrilovich and S. Markovitch (2011) Computing semantic relatedness using wikipedia-based explicit semantic analysis. Proc. of IJCAI 2007.

S. Hassan and R. Mihalcea (2011) Semantic relatedness using salient semantic analysis. Proc. of AAAI Conference 2011.

E. H. Huang, R. Socher, C. D. Manning, A. Y. Ng (2012) Improving Word Representations via Global Context and Multiple Word Prototypes. Proc. of ACL 2012.

K. Lam Tan, M. Almasri, J.-P. Chevallet, P. Mulhem, C. Berrut (2013) Multimedia Information Modeling and Retrieval (MRIM) /Laboratoire d'Informatique de Grenoble (LIG) at CHiC2013. CLEF (Working Notes) 2013.

G. Kordopatis-Zilos, G. Orfanidis, S. Papadopoulos, Y. Kompatsiaris (2014) SocialSensor at MediaEval Placing Task 2014. Working notes of MediaEval 2014.

N. O'Hare, V. Murdock (2013) Modeling locations with social media. Information Retrieval 16 (1), 30-62

J. M. Ponte, W. B. Croft (1998) A language modeling approach to information retrieval. Proc. of ACM SIGIR 1998.

A. Popescu and G. Grefenstette (2011) Social media driven image retrieval. Proc. of ACM ICMR 2011.

A. Popescu (2013) CEA LIST's participation at the CLEF CHiC 2013, *Working Notes of CLEF 2013* Valencia, Spain

A. Popescu, N. Ballas. (2014). CEA LIST's participation at MediaEval 2012 Placing Task. Proceedings of MediaEval 2012 Working notes notes of MediaEval 2012.

K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch (2011) A word at a time: computing word relatedness using temporal semantic analysis. Proceedings of the 20th international conference on World Wide Web.

P. Serdyukov, V. Murdock, R. Van Zwol (2009) Placing Flickr photos on a map. Proc. Of ACM SIGIR 2009.

P. Sorg and P. Cimiano (2012) Exploiting wikipedia for cross-lingual and multilingual information retrieval. Data & Knowledge Engineering, 74:26--45, 2012

O. Van Laere, S. Schockaert, B. Dhoedt (2012) Georeferencing Flickr photos using language models at different levels of granularity: an evidence based approach. Journal of Web Semantics 16(1):17- 31, November 2012.

O. Van Laere, Steven Schockaert, Vlad Tanasescu, Chris Jones, Bart Dhoedt (2014). Georeferencing Wikipedia documents using data from social media sources. ACM Transactions on Information Systems 32(3):12, January 2014