Capturing Entity-Based Semantics Emerging from Personal Awareness Streams

A.E. Cano, S. Tucker, F. Ciravegna

Department of Computer Science,
University of Sheffield,
Sheffield, United Kingdom
{firstinitial.surname}@dcs.shef.ac.uk

Abstract. Social activity streams provide information both about the user's interests and about the way in which they engage with real world entities. Recent research has provided evidence of the presence of emergent semantics in such streams. In this work, we explore whether the online discourse of user's social activities can convey meaningful contextual information. We introduce a user-centric methodology based on tensor analysis for deriving personal vocabularies given an entity-based context. By extracting entities (e.g. location, organisation, people) from the user's stream content, we explore the data structures that emerge from the user's interrelationship with these entities. Our experimental results revealed that the simultaneous correlation of entities leads to the identification of concepts which are relevant to the user given a specific context. This methodology is relevant for mobile application designers (1) in fostering user entity-based ontologies for merging user context in pervasive environments, (2) for personalising entity-based recommendations.
Keywords: linked data streams, social awareness streams, microblogging, context

1 Introduction

The past few years have seen the launch of different social networking platforms that allows a user to expose their online presence, create groups and build bridges for communicating within their online social spheres. The high usage of these platforms has generated an enormous amount of personal information online creating unprecedented opportunities for a wide range of research related to knowledge management, user contextualisation, and the Semantic Web.

In this paper, we focus on the analysis of a user's social activity streams (a.k.a personal awareness streams [32]) generated from different social networks.We consider a user's social activity stream as a historical dataset from which context-sensitive items can be derived. Users produce data streams, not only providing information regarding the physical world (e.g. location, surrounding things) but also regarding their digital environment (e.g. adding new friends, microblogging). Therefore, we see the user's social activity streams as virtual sensors that could provide valuable information not only about the user interests but also about the user's physical contextual situation.

This paper sets out to explore whether the use of aggregations of personal awareness streams can convey meaningful contextual information given a set of different

entities that the user has interacted with within their online discourse during a timeline. In this paper, we introduce the Concept Selection Induced from Social Stream Aggregations (CSISSA) methodology, which captures entity-related information (e.g. organisations, locations, people, links) emerging from a personal awareness stream aggregation. This methodology is based on a three-mode network of social awareness streams (a.k.a. Tweetonomy [32]) and lightweight associative resource ontologies [20]. CSISSA applies tensor analysis for performing a simultaneous correlation of the given entities. Computing the decomposition of the tensor yields to conceptual structures that characterise a user given a context.

In this work we investigate the way in which a user refers to entities in the content of the message he generates. These entities are interlinked to others through, for example text and hashtags. We explore if this entity-based interrelationship can yield emerging conceptual structures that can aid in the user modelling. Our experimental results suggest that a key factor for successfully deriving relevant concepts for a given context is the user's microblogging verbosity, and the use of common vocabularies referring to the entities involved in the context.

The contributions of this paper are as follows: we study personal awareness stream aggregations as a source of information for deriving users' relevant concepts given an entity-based context. We present a novel approach which enables the explicit declaration of the context in which a user needs to be analysed. Our model abstracts the semantics of the vocabularies introduced by the user in his social activity stream by means of the derivation of lightweight ontologies. We make use of tensor analysis for building a user's entity-based context.The encapsulation of an entity-related lightweight ontology constitutes a slice of a tensor. The decomposition of this tensor reveals concepts relevant to the user in the analysed context. We believe that entity-based user modelling could aid in the future integration of user context to pervasive environments.

2 Background

In this section we start by defining concepts from principal component analysis (PCA) and then we give a brief introduction to tensor analysis. We will follow the typical conventions, and denote matrices with upper case bold letters (e.g. $\mathbf{X}$, row vectors with lower-case letters (e.g. $\mathbf{v}$), and tensors with calligraphic font (e.g., $\mathcal{X}$).

Principal Component Analysis (PCA) PCA [8] helps to identify patterns in data by expressing this data in such a way that it highlights a limited number of "components" that capture most of the information contained in the observed variables. By performing an orthogonal linear transformation, PCA finds the best linear projections which minimize least squares cost. For a given matrix $\mathbf{X}$ with zero mean (i.e. the mean of the distribution has been subtracted from the data set), PCA can be computed by obtaining the Singular Value Decomposition (SVD)[8][2] of $\mathbf{X}$; according to which $\mathbf{X} = \mathbf{U}_{svd} \times \mathbf{\Sigma}_{svd} \times \mathbf{V}^T_{svd}$; then $\mathbf{Y} = \mathbf{U}_{svd} \times \mathbf{\Sigma}_{svd}$ and $\mathbf{U} = \mathbf{V}_{svd}$. For example, if $\mathbf{X}$ is a user's status-keywords matrix taken from a user's stream aggregation dataset, then the $\mathbf{Y}$ and $\mathbf{U}$ matrices can be interpreted as the status-concept matrix $\mathbf{Y}$, and the keywords-concept matrix $\mathbf{U}$.

A user's post can be further analysed by considering not only keywords but also other resources (e.g. location, people) embedded on its content; forming a multidimensional set of parameters. An example for such analysis could study those topics that emerge from a user's posts generated during the morning hours at the office (location×time×keywords). A mathematical abstraction for the representation of a higher way structured data is a Tensor.

Tensor Analysis Tensors[12] are multidimensional M-ways or Mth-order arrays which generalize the notion of vectors(1-way or first-order array) and matrices (2-ways or second-order arrays). Tensors of order greater or equal to three are called higher-order tensors. In order to identify patterns that emerge from the simultaneous correlation of a set dimensions it is necessary to decompose a tensor. Tensor decomposition can be considered as a higher-order generalisation of SVD and PCA. In this paper we will use the Tucker decomposition approach.

Tucker Decomposition The Tucker decomposition was first introduced by Tucker in 1963 [30]. Given a tensor $\mathcal{X} \in R^{I_1 \times \ldots \times I_N}$ PCA is performed so as to decompose tensor $\mathcal{X}$ into a core tensor $\mathcal{G} \in R^{R_1 \times .. \times R_N}$ multiplied by a set of matrices $\mathbf{U}^{(i)} \in R^{I_i \times R_i}$. Therefore the Tucker decomposition of a three-order tensor $\mathcal{X}$ can be expressed as.

$\mathbf{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} =$
$\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} \mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r \equiv [[\mathcal{G}; \mathbf{A}, \mathbf{B}, \mathbf{C}]]$

One of the approaches for computing a Tucker decomposition of a three-order vector is to start with a first approximation obtained by applying a Higher Order SVD (HOSVD) [16] and then apply the alternating least squares algorithm (ALS) [15].

3 Related Work

Mika [20][28] explores how community-based semantics, in the form of lightweight associative ontologies, emerge from folksonomies. He introduces the semantic-social networks model which consists of a tripartite graph of people, concept and instance associations. Wagner and Strohmaier [32] introduce the Tweetonomy model, which is a formalisation of social awareness streams. This model adopts a theoretic approach similar to the one presented by Mika. However, the Tweetonomy model presents a more complex and dynamic structure than folksonomies. Strohmaier et al[18] and Körner et al[13], study quantitative measures for tagging motivation. In their study they found empirical evidence that the emerging semantics of tags in folksonomies are influenced by individual user tagging practices.

Tensor decompositions have a long history and have been applied in different research communities. In particular the Tucker decomposition has been used in chemical analysis [4], psychometrics [9] and computer vision [31]. Tensor analysis has also been applied in web search; Kolda et al [11] propose a method called Topical HITS (TOPHITS) which can be considered as an extension of Kleinberg's HITS (Hypertext Induced Topic Selections) algorithm [10]. TOPHITS analyses a semantic graph that combines anchor text with the hyperlink structure of the web. In order to avoid losing edge type information when modelling the adjacency structure of a semantic graph as a matrix, they modelled it as a three-way tensor containing both the hyperlink and an-

chor text information. Their tensor decomposition leads to triplets of vectors containing authority, hub scores for the pages, and topic scores for the terms.

Rendle and Thieme [25] apply tensor factorisation for personalised tag recommendation and learning. They introduce a model based on Tucker decomposition to explicitly model the pairwise interaction between users, items and tags. More similar to our work is the approach of Wetzker et al [33]. They follow a user-centric tag model for deriving mappings between personal tag vocabularies (a.k.a personomies [6]) and the corresponding folksonomies. Our approach differs from previous work in that rather than building the tensor as a three-way tensor of items-users-tags, we generate a three-way tensor in which each slice is a lightweight associative "resource" ontology; which allows to store multiple stream qualifiers in the tensor.

The analysis of user-generated content extracted from social media sites is an active research area. Qualitative and quantitative studies have been carried out for leveraging the "wisdom of crowds" [22]. Some of this research has focused on questions related to network and community structure. For example, Krishnamurthy et al [14] present a characterisation of Twitter social network, which includes patterns in geographic growth and user's social activity. In their work, they suggest that frequent updates might be correlated with high overlap between friends and followers. Java et al [7], present an analysis of Twitter and suggest that the differences in users' network connection structures can be explained by the following types of user activities: information seeking, information sharing and social activity.

Other work has presented a systematic analysis of the content of posts in social networks. Recent work [21], introduces the term "Social Awareness Streams" for referring to this aggregation of short status messages. They proposed a characterisation of these messages via a human coding of tweets into nine categories including "Information sharing" and "Self promotion". By extrapolating from these categories, they induced two types of users the "informers", who post about non-personal information, and the "meformers" which mostly post about themselves. Stankovic et al [17], study conference related tweets. They map tweets to talks an subevents that they refer to. Using linked data they derive additional knowledge about event dynamics and user activities.

Data structures emerging from the Social Web have been studied in the Information Retrieval and Semantic Web communities. Research in this area includes the study of content and link analysis algorithms and ontology learning algorithms. Heymann et al [5] present an algorithm for hierarchical taxonomy generation from social tagging systems. For generating a taxonomy of tags, they apply graph centrality in a cosine similarity graph of tags. Ramage et al [23], apply labelled Latent Dirichlet Allocation (LDA) [24] for mapping content of the Twitter feed into four dimensions including style and substance. Schmitz [26] introduces a subsumption-based model for inducing faceted ontologies from Flickr tag vocabulary. Our work was inspired mainly by Mika's [20], and Wagner and Strohmaier's [32] work. We apply the Tweetonomy formalisation for obtaining personal awareness stream aggregations. Our work differs from existing work (1) through our focus on deriving person-based lightweight ontologies from personal awareness stream; which enrich concepts and reveal structures that are meaningful to the owner of the stream; (2) we study the content of the messages not only in terms of traditional resources as hashtags, and links, but also in terms of entities (e.g loca-

tion, people, organisations); (3) we present a methodology based on tensor analysis that allows the definition of entity-based context for deriving person-based ontologies.

4 Social Stream Aggregation and Entity-Based Concept Induction

Our interest is to enable a way in which a user's social activity streams can be analysed in order to discover concepts that can aid in profiling him. These concepts are revealed as a combination of featuring dimensions. Example of these dimensions include e.g. a user's interests, user location, user's tendencies in favouring a position in a discussion etc. The following subsection presents the definition of three different social networks modelled as tripartite social awareness streams.

4.1 User's Social Stream Aggregation

Following the Tweetonomy model suggested by Wagner and Strohmaier[32], we describe a social awareness stream as a sequence of tuples S, according to the following definition:

Definition 1. *A tweetonomy is a tuple* $\mathtt{S} := (\mathtt{U}_{\mathtt{q1}}, \mathtt{M}_{q2}, \mathtt{R}_{q3}, \mathtt{T}, ft)$*, where*

- $\mathtt{U}$,$\mathtt{M}$,$\mathtt{R}$ *are finite sets whose elements are called users, messages and resources.*
- *Each of these sets are qualified by* $\mathtt{q1}$,$\mathtt{q2}$*, and* $\mathtt{q3}$ *respectively (explained below).*
- $\mathtt{T}$ *is the ternary relation* $\mathtt{T} \subseteq \mathtt{U} \times \mathtt{M} \times \mathtt{R}$ *representing a hypergraph with ternary edges. The hypergraph of a tweetonomy* $\mathtt{T}$ *is defined as a tripartite graph* $\mathtt{H}(\mathtt{T}) = \langle \mathtt{V}, \mathtt{E} \rangle$ *where the vertices are* $\mathtt{V} = \mathtt{U} \cup \mathtt{M} \cup \mathtt{R}$*, and the edges are:* $\mathtt{E} = \{\{u, m, r\} \,|\, (u, m, r) \in T\}$*. Each edge represents the fact that a given user associates a certain message with a certain resource.*
- f_t *is a function that assigns a temporal marker to each ternary edge.*

In this study we will focus on user-centric social streams generated in Facebook, Foursquare and Twitter, according to the following qualifiers:

- The way a user can be related to a message is represented by the qualifier $\mathtt{q1}$. For this analysis we only consider the authorship relationship: $\mathtt{U}_a$ (the author of the message).
- The qualifier $\mathtt{q2}$ represents the types of messages. This is a comment or a status in Facebook; a broadcast message, direct message, re-tweeted message in Twitter; a broadcast message (shout) in Foursquare are considered to be the same type. For this experiment we don't differentiate between types.
- The qualifier $\mathtt{q3}$ for resources considers: $\mathtt{R}_k$ (keywords), $\mathtt{R}_h$ (hashtags), $\mathtt{R}_{li}$ (URLs), $\mathtt{R}_{mlo}$ (message-emitted location), $\mathtt{R}_o$ (organisations - entities recognised as an organisation), $\mathtt{R}_p$ (people -entities recognised as a person), $\mathtt{R}_l$ (location - entities recognised as a location).

We focus on a user given the streams he has produced within a window of time. Given the tuples $\mathtt{T}_{\mathtt{facebook}}, \mathtt{T}_{\mathtt{foursquare}}, \mathtt{T}_{\mathtt{twitter}}$, we define the sets $\mathtt{U}, \mathtt{R}, \mathtt{M}$ as:

$\mathrm{U} = \mathrm{U}_{\mathtt{facebook}} \cup \mathrm{U}_{\mathtt{twitter}} \cup \mathrm{U}_{\mathtt{foursquare}}$,

$\mathrm{R} = \mathrm{R}_{\mathtt{facebook}} \cup \mathrm{R}_{\mathtt{twitter}} \cup \mathrm{R}_{\mathtt{foursquare}}$,

$\mathrm{M} = \mathrm{M}_{\mathtt{facebook}} \cup \mathrm{M}_{\mathtt{twitter}} \cup \mathrm{M}_{\mathtt{foursquare}}$

We are interested in extracting the concepts emerging from the streams produced by a user:

$\breve{\mathtt{u}} \in \mathtt{U} : \breve{\mathtt{u}} \in \mathrm{U}_{\mathtt{facebook}} \wedge \breve{u} \in \mathrm{U}_{\mathtt{twitter}} \wedge \breve{\mathtt{u}} \in \mathrm{U}_{\mathtt{foursquare}}$

In order to do so we consider a user stream aggregation defined as a tuple:

$\mathtt{S}_{\mathtt{a}}(\mathtt{U}') = (\mathtt{U}, \mathtt{M}, \mathtt{R}, \mathtt{Y}', ft)$, where

$\mathtt{Y}' = \{(\mathtt{u}, \mathtt{m}, \mathtt{r}) \mid \mathtt{u} \in \mathtt{U}' \vee \exists \mathtt{u}' \in \mathtt{U}', \tilde{\mathtt{m}} \in \mathtt{M}, \mathtt{r} \in \mathtt{R} : (\mathtt{u}', \tilde{\mathtt{m}}, \mathtt{r}) \in \mathtt{Y}\}$

and $\mathtt{U}' \subseteq \mathtt{U}$ and $\mathtt{Y}' \subseteq \mathtt{Y}$. $\mathtt{S}_{\mathtt{a}}(\mathtt{U}')$, consists of all messages related with a user $\mathtt{u}' \in \mathtt{U}'$ and all the resources and users related with these messages.

4.2 Lightweight Associative Ontologies

An ontology, is a shared, formal conceptualization of a domain [3][1]. It is a data structure which is an advancement in conceptual modelling over taxonomic structures [28]. A lightweight ontology can be considered as an evolving classification structure created by users [27], which can be considered to be closer to a thesaurus (i.e. a structure organising topics). We want to derive a set of concepts from a simultaneous correlation among the resources q3 (e.g. keywords, hashtags, links) extracted from a user stream aggregation. In order to obtain this correlation, we start identifying those bipartite graphs (two-mode graphs) that could be of any interest to our analysis.

Consider for instance the association between keywords and location; which can be obtained as a combination of location×message ($\mathbf{R}_l\mathbf{M}$) and keywords×messages ($\mathbf{R}_k\mathbf{M}$). Where the location×messages (bipartite graph $\mathbf{R}_l\mathbf{M}$) is defined as:

$\mathrm{R}_l\mathbf{M} = \langle \mathtt{R_l} \times \mathtt{M}, \mathtt{E_{rm}} \rangle = \{(r, m) \,|r \in \mathtt{R_l} \wedge \exists u \in U : (u, m, r) \in E\}$,

$w : E \rightarrow R, \forall e = (r, m) \in \mathtt{E_{rm}}$

and the keywords×message (bipartite graph $\mathbf{R}_k\mathbf{M}$), is defined as:

$\mathrm{R}_k\mathbf{M} = \langle \mathtt{R_k} \times \mathtt{M}, \mathtt{E_{rm}} \rangle = \{(r, m) \,|r \in \mathtt{R_k} \wedge \exists u \in U : (u, m, r) \in E\}$,

$w : E \rightarrow R, \forall e = (r, m) \in \mathtt{E_{rm}}$

These bipartite graphs represent the adjacency or affiliation matrices: $\mathbf{R}_l\mathbf{M}$; which links the resources (of type location) to the messages in which this resource has been mentioned by this user. In the same way, $\mathbf{R}_k\mathbf{M}$; links the resources (of type keyword) to the messages in which this resource has been mentioned by at least one user. Each link (edge) can be weighted following a local or global weighting function in order to condition the data to be analysed (see Fig. 1).

Finally, the association between keywords and location is expressed as $\mathbf{R}_k\mathbf{R}_l = (\mathbf{R}_k\mathbf{M})(\mathbf{R}_l\mathbf{M})^{\mathbf{T}}$. We can now encapsulate the information that associates locations with keywords only in terms of keywords by multiplying $\mathbf{R}_k\mathbf{R}_l$ with its transpose, i.e. $\mathbf{O}(\mathbf{R_k}\mathbf{R_l}) = (\mathbf{R_k}\mathbf{R_l})(\mathbf{R_k}\mathbf{R_l})^{\mathbf{T}}$. This matrix, known as co-affiliation matrix, can be considered as a lightweight associative location ontology [20] based on overlapping sets of keywords.

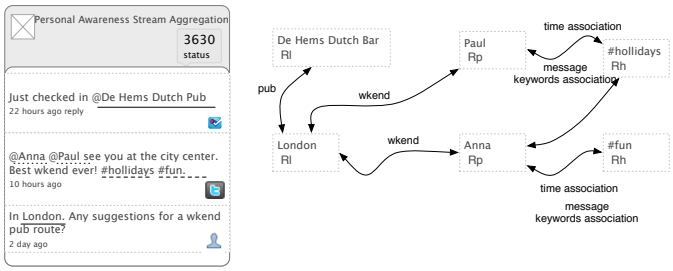

Fig. 1. A personal awareness stream on the left yields the semantic graph on the right, formed of resources of type location, people and hashtags. The edges in the graph are labelled with the resources that link the entities.

4.3 Concept Selection Induced from Social Stream Aggregations (CSISSA)

In this paper we propose the **Concept Selection Induced from Social Stream Aggregations** technique. This technique obtains a set of concepts derived from the simultaneous analysis of the correlation of different stream qualifiers. It is based on the analysis of Sp3way tensors [29] in which each slice consists of a dense matrix formed by the product of a sparse matrix and its transpose. The motivation for using this class of tensors arises from the need of simultaneously storing multiple stream qualifier matrices.

Given P lightweight ontologies characterising a user's social streams consisting of N messages; we define a tensor $\mathcal{O} \in R^{N \times N \times P}$ consisting of frontal slices of the form $\mathbf{O}_p = \mathbf{B}_p \mathbf{B}_p^T$ with $p = 1, ..P$, where $\mathbf{B}$ is a bipartite graph deriving the lightweight ontology $\mathbf{O}_p$; see Figure 2.

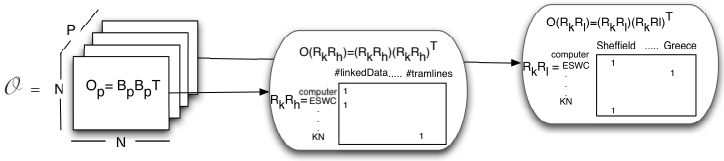

Fig. 2. Lightweight ontology tensor O.

The computation of a Tucker decomposition (presented in subsection 2) of $\mathcal{O}$ yields to an approximation of the form

$\mathrm{O} \approx \mathcal{G} \times_1 \mathbf{K} \times_2 \mathbf{K}' \times_3 \mathbf{C} =$

$\sum_{i=1}^{N} \sum_{i=1}^{N} \sum_{p=1}^{P} g_{iip} \mathbf{m}_i \circ \mathbf{m}'_i \circ \mathbf{c}_p \equiv [[\mathcal{G}; \mathbf{K}, \mathbf{K}', \mathbf{C}]]$

The output has the property $\mathbf{K} \approx \mathbf{K}'$, the rows of these matrices contain feature vectors that encapsulate a compilation of the different similarities expressed in the frontal matrices. $\mathbf{K}$ and $\mathbf{K}'$, can be regarded as keyword $\times$ keyword-group matrices highlighting those keywords that are more relevant to the similarities expressed in all $\mathbf{O_p}$. The

matrix $\mathbf{C}$ represents an index $\times$ index-group matrix which highlights $\mathbf{O_p}$ matrices. Finally the tensor $\mathcal{G}$ expresses how groups (keywords-group and index-group) relate to each other. The frontal matrix K highlights those concepts.

5 Deriving Relevant Concepts with CSISSA

The analysis with CSISSA is carried out on a user's social stream aggregation $\mathtt{S_a}(\mathtt{U}')$, this aggregation is built upon the messages the user has posted in different social networks. These messages are saved in a data store as the user generates them, and can be retrieved in windows of time of n days, this is: $\mathtt{S_a}(\mathtt{U}')\,[t_s, t_e] = (\mathtt{U}, \mathtt{M}, \mathtt{R}, \mathtt{Y}', f_t)$, where $f_t : \mathtt{Y}' \rightarrow N, t_s \leq f_t \leq t_e$ and $|t_e - t_s| = n$ days.

The retrieved messages need to be pre-processed; 1) Stop words, punctuation and numbers from the message content are removed; 2) From the message content, entities of type: Location, Person and Organisation are extracted. Qualifiers of type: keywords, hashtags and geocodes (when provided) are also extracted. This section presents a concrete example in which CSISSA can be applied.

5.1 Recurrent Entity-Concept Analysis

Consider the problem of finding a temporal correlation among certain entities to which a user is engaged with, through the messages he has posted within a window of time; and from these entities induce a set of concepts to which they can be linked (this can be applied in temporal user profiling and event detection). The selection of the correct bipartite graphs to take part on the three-order tensor depends on the situation from which the entity-based context needs to be extracted. For example, considering the entities: Hashtag and Location; we define the following lightweight ontologies:

- **Lightweight Associative Keyword Ontology** Given a keyword $\times$ message matrix $\mathbf{R_kM} = w_{ij}$, where w_{ij} is computed following a term frequency-inverse document frequency (tf-idf) weighting function [19]. We define the lightweight associative keyword ontology $\mathbf{O}(\mathbf{R_kM})$ as $\mathbf{O}(\mathbf{R_kM}) = (\mathbf{R_kM})(\mathbf{R_kM})^\mathbf{T}$.
- **Lightweight Associative Hashtag Ontology**, we define the hashtag $\times$ message matrix $\mathbf{R_pM}$ following as well a (tf-idf) weighting function. The $\mathbf{O}(\mathbf{R_hM})$ is defined as $\mathbf{O}(\mathbf{R_hM}) = (\mathbf{R_hM})(\mathbf{R_hM})^\mathbf{T}$.
- **Lightweight Associative Location Ontology**, we define the places $\times$ message matrix $\mathbf{R_lM}$ following as well a (tf-idf) weighting function. The $\mathbf{O}(\mathbf{R_lM})$ is defined as $\mathbf{O}(\mathbf{R_lM}) = (\mathbf{R_lM})(\mathbf{R_lM})^\mathbf{T}$.
- **Ligthweight associative time ontology**, first, we obtain the hour $\times$ message affiliation matrix $\mathbf{HM} = v_{ij}$ where $v_{ij} = 1$ if the time message m_j was produced during the hour h_i and $v_{ij} = 0$ otherwise. We define the ligthweight associative time ontology $\mathbf{O}(\mathbf{HM})$ as $\mathbf{O}(\mathbf{HM}) = (\mathbf{HM})(\mathbf{HM})^\mathbf{T}$.

To analyse the correlation of these entities and derive the related concepts, it is necessary to encapsulate the previous ontologies in terms of keywords (see section 4.2); i.e to obtain $\mathbf{O}(\mathbf{R_kR_h})$, $\mathbf{O}(\mathbf{R_kR_l})$, $\mathbf{O}(\mathbf{R_kH})$. These ontologies will form the slices of

the tensor $\mathcal{O}$. The computation of a Tucker decomposition of the $\mathcal{O}$ tensor will reveal a ranked vector of concepts. By decomposing each of the tensor slices, it is possible to derive the entities relevant to the decomposition.

Table 1, presents the relevant concepts, and the highlighted entities derived from the Tucker decomposition of a tensor built from the stream aggregation of one of the users we followed in our evaluation (see section 6). This analysis reveals concepts that are recurrently relevant to the user. In this case, these results expose the correlation of the locations: Sheffield, London and Washington with the user's work related concepts during working hours.

Table 1. Concepts in the context of Hashtags-Places-Time

Emerged Concepts	linkeddata, semanticweb, talis, data.ac.uk, wrt, link, quality, astonbusinessschool, environment, funded
Hash tags	#linkeddata, #semanticweb, #talis, #astonbusinessschool, #linkquality, #ldal, #sheffield, #isko, #informationextraction, #unsupervisedclustering
Places	London, Sheffield, Washington
Time	[9:00am-5:00pm], [7:00pm-11:00pm]

6 Evaluation and Conclusions

CSISSA was evaluated on the grounds of the relevance of a concept induced by a given contextual need. A contextual need was expressed by a pair of contexts, e.g. Location-Time, Hashtag-Location. CSISSA provides a set of relevant concepts computed by the simultaneous correlation of the entities involved in a given context. For testing this technique, we "followed" a set of four "active" microbloggers. Three of them technology oriented user, and one of them an active blogger in education. The stream aggregations were recorded from 1st of July until the 25th September 2010 , and entities where extracted using Open Calais services [1].

In the absence of a gold standard, evaluating the concepts that emerge from a user's social aggregation given a context is a difficult task; it requires consulting the author of the social stream whose context-induced concepts are being mapped. For evaluating the effectiveness of CSISSA, each user was presented with a contextual need, and a set of concepts derived by CSISSA. The users were asked to mark each concept as relevant or irrelevant to the given context. Although CSISSA allows the simultaneous correlation of n-entities, which define the context; we performed the evaluation on a maximum of two entities at a time. The evaluated contexts are: hashtag-time, location-people, and organisation-people. For example, by deriving concepts related to hashtag and time for one of the users, the question was: *In terms of the association between the hashtag #linkeddata, and the timeslots ([12pm-5pm], 8pm), which of the following concepts do*

[1] Open Calais, http://www.opencalais.com/

you consider relevant?. For the hashtag-time context, three different hashtags where evaluated, and in the same way for the other two contexts.

As it is well known, acquiring the relevance judgement of all the ranked concepts in terms of precision/recall is a time-consuming and expensive process [19]. Mainly because the ranked vector can consist of hundreds of concepts that a user would not be willing to evaluate. Therefore, we have decided to use the Mean Average Precision (MAP) metric [19]. MAP measures the mean of the precision scores obtained after each relevant concept is retrieved, using zero as the precision for relevant concepts that are not retrieved. The MAP value represents the average under the precision-recall curve for a set of queries. MAP values were averaged for the three cases of each context. The results are depicted in Figure 3 a), which shows a generalized MAP performance of the relevancy of the concepts judged by each user given a context using CSISSA.

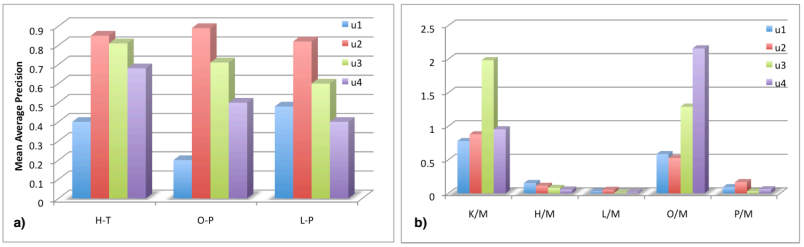

Fig. 3. a) Mean average precision (MAP) performance by user and contextual information need including HashTag-Time, Organisation-People, and Location-People, for the top 15 concepts. b) Normalised Lexical (Number of keywords(K)/ Number of Messages(M)), Topical (Hashtag), Spatial, Organisation-Entity, People-Entity Diversity.

These results suggest that higher lexical diversity (K/M) leads to better MAP results (see Figure 3 b)), this is an expected result since CSISSA explores the way in which an entity is linked to another one through keywords. We expected to discovered relevant concepts first if the user exposed a correlation between contexts, and second if this correlation was able to be expressed by keywords.

However, although the microblogging verbosity provided a better basis for deriving meaningful concepts, the relevance of the concepts given a context depended highly on the user's patterns of correlating the entities through keywords. In our experiments a fairly naive approach was taken by not considering the ambiguity in which user's can relate two entities with a keyword. Future work considers the introduction of concept disambiguation for tackling this issue.

CSISSA enabled to model users' generated patterns in their social activity streams given an entity-based context. These patterns expose the implicit association in which the user interlinks entities. The concepts derived with CSISSA suggests their applicability in user modelling, and the awareness of user intentions. A main implication of our work is that personal awareness streams can be used effectively to model context

by leveraging the user's entity affiliations. We believe that our approach can also help in merging user contexts in pervasive environments.

During the evaluation, one of the users did not remember to have tweeted about a particular topic, until we showed him the tweet, this suggest the necessity of introducing relevance-decay functions in our calculations. We also noticed that many of the users' streaming topics' relevance was in many cases volatile; further research is necessary to address these issues. We are also planning to test this technique on a bigger corpus, and to compare this technique against other baselines e.g. topic analysis.

7 Acknowledgements

This work has been supported by the European Commission as part of the WeKnowIt project (FP7-215453), and partially supported by CONACyT, grant 175203

References

1. Borst, Pim, Akkermans, Hans, and Top, Jan. Engineering ontologies. *International Journal of Human-Computer Studies*, 46(2-3):365–406, February 1997.
2. G. H. Golub and C. F. Van Loan. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*. The Johns Hopkins University Press, 3rd edition, October 1996.
3. Gruber, Thomas R. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June 1993.
4. R. Henrion. N-way principal component analysis theory, algorithms and applications. *Chemometrics and Intelligent Laboratory Systems*, 25(1):1–23, 1994.
5. P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford University, April 2006.
6. A. Hotho, R. JÃschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*, chapter 31, pages 411–426–426. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2006.
7. A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.
8. I. T. Jolliffe. *Principal Component Analysis*. Springer, second edition, October 2002.
9. V. M. I. Kiers HA. Three-way component analysis: principles and illustrative application. *Psychol Methods*, 6(1):84–110, 2001.
10. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:668–677, 1999.
11. T. Kolda and B. Bader. The TOPHITS model for higher-order web link analysis. In *Proceedings of the SIAM Data Mining Conference Workshop on Link Analysis, Counterterrorism and Security*, 2006.
12. T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.
13. C. Körner, D. Benz, A. Hotho, M. Strohmaier, and G. Stumme. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 521–530, New York, NY, USA, 2010. ACM.

14. B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM.
15. P. M. Kroonenberg and J. D. Leeuw. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45:6997, 1980.
16. B. D. M. L. De Lathauwer and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21:12531278, 2000.
17. M. R. M. Stankovic and P. Laublet. Mapping tweets to conference talks: A goldmine for semantics. In *Proceedings of Social Data on the Web workshop, ISWC 2010. Shanghai, China*. ISWC 2010, 2010.
18. R. K. M. Strohmaier, C. Körner. Why do users tag? detecting users' motivation for tagging in social tagging systems. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM2010)*, Washington, DC, USA, 2010.
19. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
20. P. Mika. Ontologies are us: A united model of social networks and semantics. In *In Proceedings of ISWC 2005*, 2005.
21. M. Naaman, J. Boase, and C. H. Lai. Is it really about me?: message content in social awareness streams. In *CSCW '10: Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192, New York, NY, USA, 2010. ACM.
22. O'Reilly, Tim. O'Reilly Network: What Is Web 2.0, September 2005.
23. D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.
24. D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Morristown, NJ, USA, 2009. Association for Computational Linguistics.
25. S. Rendle and L. S. Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 81–90, New York, NY, USA, 2010. ACM.
26. P. Schmitz. Inducing ontology from flickr tags. In *WWW 2006, May 22-26, 2006, Edinburgh, UK*. IW3C2, 2006.
27. L. Specia and E. Motta. Integrating folksonomies with the semantic web. In *In Proc. of the 4th ESWC*, pages 624–639, 2007.
28. Technology, Knowledge, P. Mika, and H. Akkermans. Towards a new synthesis of ontology. Technical report, 2004.
29. W. K. T.M. Selee, T. Kolda and J. D. Griffin. Extracting clusters from large datasets with multiple similarity measures using imscand. In *CSRI Summer Proceedings*, 2007.
30. L. R. Tucker. *Implications of factor analysis of three-way matrices for measurement of change*. C. W. Harris, University of Wisconsin Press, 1963.
31. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *In Proceedings of the European Conference on Computer Vision*, volume 1, pages 447–460, 2002.
32. C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*, april 2010.
33. R. Wetzker, C. Zimmermann, C. Bauckhage, and S. Albayrak. I tag, you tag: Translating tags for advanced user models. In *WSDM '10: Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 71–80, New York, NY, USA, 2010. ACM.