Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

D1.2 Data sets from folksonomic sites

Period covered: from 01/06/2007 to 31/05/2008
Start date of project: June 1st, 2006
Due date of deliverable: July 15th, 2008
Distribution: Public

Date of preparation: 31/05/2008
Duration: 36 months
Actual submission date: June 20th, 2008
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: “Sapienza” Università di Roma
Lead contractor for these deliverables: “TAGora partners”

Contents

Chapter 1

Data collection from bibliographic reference sharing system

1.1 Introduction

The TAGora consortium remains fully committed to the release of data and software whenever it is possible and beneficial to do so. To this end, the project web site has been extended with an additional page[1] that is specifically dedicated for the data collected and/or used by the project.

Currently, the data collections web page contains links for downloading the various datasets, such as Netflix, IMDB, Bibsonomy, MOTOBOY, UK and Last.fm music singles charts, IKoru, etc. The data that will be withheld for the time being include most data from Flickr, Last.fm, and del.icio.us. These datasets will be made available later in the project.

This document lists the data sets currently available to the TAGora consortium. For each data set, we provide a description of its content, data type and quantity, format, and links to where the data can be downloaded from if the data is *public* or has been *anonymized*. We also point to some of the TAGora publications where the individual data sets have been used.

1.2 Integrated IMDB and Netflix Dataset

To support the investigation of communal data structures, such as folksonomies, in the context of recommendation, we have created a large knowledge base about movies and how users rate movies. To achieve this, a large portion of the Internet Movie Database (IMDB) was downloaded[2] to provide information about movies, actors and production personnel, as well a large set of keywords that have been assigned by users to describe movies. The IMDB dataset contains 898,078 movie titles, 2,564,990 names (including actors, actresses, writers, directors and producers), and 32,247 keywords. To obtain information about the way users rate movies, we have collected a dataset[3] from Netflix, a mail-based movie rental company in the US, which contains the movie ratings of 480,189 customers across 17,770 movie titles over the last five years.

Both the IMDB and Netflix datasets have been converted into a relational database, a 643MB compressed MySQL dump. To provide a single view over both datasets, for example, to support the querying of information on movies from IMDB and how users rate these movies from Netflix, we have correlated the 13,880 movie titles in the Netflix dataset with their IMDB counterparts. The result is a large knowledge base on movies and movie ratings that supports semantic querying (for example through SPARQL). The mappings between movie titles in Netflix with those

[1] `http://www.tagora-project.eu/data/`

[2] `http://www.imdb.com/interfaces`

[3] `http://www.netflixprize.com/download`

in IMDB can be downloaded from http://users.ecs.soton.ac.uk/mns2/data/netflix_imdb_mapping.csv.

This dataset was used to study movie recommendations (Szomszor et al., 2007).

1.3 Flickr photos

This data collection contains all descriptions of photos that were uploaded to Flickr during January 2004 and December 2005 and that were still available over the public API in the first half of 2007. The crawling of the data collection was finished 07/2007.

The collection contains information about 320K users, 28M photos, 1.6M tags and 113M tag assignments. Parts of the collection were used in the papers (Abbasi and Staab, 2008) and (Görlitz et al., 2008). The data collection is available in an anonymized form https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/PINTSExperimentsDataSets.

1.4 Flickr user data and group memberships

This data collection contains information about 3M Flickr users, i.e. their contact lists, favorite photos and memberships in Flickr user groups. The crawl was done via the public API during 08/2007.

The collection contains information about 3M users, 29M contacts on their contact lists, 5.5M tags, 41M photos, 187M tag assignments 77M comments and 3M notes attached by the users to the photos, 50M photos on their favorite lists. Additionally, 132K different user groups of Flickr were crawled that include information about 13M memberships in the groups.

The data collection will mainly be used for research about recommender systems and is not yet publicly available.

Another dataset of Flickr was gathered for the IKoro application. The dataset consists of data about 12 users, 3911 unique tags, 3300 photos, and 25157 tag assignments. This data is stored in a MySQL database, and was collected in the autumn of 2005.

1.5 Del.icio.us

Data from del.icio.us was gathered in 2006 and currently consists of over 667 thousand users, nearly 2.5 million tags (organized in 667 bundles), and around 18.7 million resources.

This data set was extensively used in the project, for example in the analysis and modelling of evolutional behaviour and structural information of social resource sharing systems, analysis and modelling of the structure and dynamics of folksonomies, and in semantic user interest profiling and tag disambiguation analysis.

Some of the work that used this dataset is described in the papers: (Cantador et al., 2008; Cattuto et al., 2007; Grahl et al., 2007; Hotho et al., 2006; Szomszor et al., 2008b).

1.6 Ajax, Blog and XML co-occurrence streams

For the paper (Dellschaft and Staab, 2008), we created a sub-collection of the larger Del.icio.us crawl. The collection consists of the complete co-occurrence streams of the “ajax”, “blog” and “xml” tag. The data set and more detailed information about it (e.g. its size) are available at https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/Tagdataset.

1.7 Search sessions of the MyTag portal

This data collection contains the anonymous search sessions that were submitted via the MyTag portal (http://mytag.uni-koblenz.de). It contains information about the searched tags and platforms. Furthermore, the resources that were selected from the result lists are given. Subsequent search requests of a single user are connected via a session ID. The collection is created via a database dump of the MyTag portal and will be regularly updated. The data set and more detailed information about it (e.g. its size) are available at https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/MyTag.

1.8 Bibsonomy

To provide the Consortium with raw data for modeling and analyzing interactions in online social communities, we offer a benchmark dataset from our collaborative tagging system BibSonomy. The anonymized data of BibSonomy are downloadable via a MySQL dump, which will be updated every half year. Interested people get an account from kde@cs.uni-kassel.de for access to our server on https://www.kde.cs.uni-kassel.de/bibsonomy/dumps. Before starting the download, participants have to sign a license agreement in which terms of use are set up. The dump from December 2006, for instance, includes data from approximately 400 users, 12.000(different)/140.000(all) tags and 39.000 resources. The dumps can easily be loaded into a MySQL database.

1.9 MOTOBOY

The dataset from the canal*MOTOBOY project[4], which involves a small-scale community using tags to represent and communicate their daily life experiences has been made available to the TAGora consor tium. In canal*MOTOBOY, 15 motorcycle messengers in Sao Paulo Brazil transmit tagged images, videos and audio clips directly from their mobile phones to a web page. The dataset, which includes 13 months of activity, can be used to study the dynamics of tagging of a small, densely-connected group. It contains over 8000 tag assignments, nearly 8000 esources, 712 tags and 15 users.

This dataset can be downloaded from http://www.csl.sony.fr/~tisselli/zexe/zexe_motoboy.csv.

1.10 Last.fm

To study music tags, we collected data from Last.fm about users, tags, artists, albums, tracks, sound extracts. The data is currently stored in a MySQL database, and consists of 10 users, 200 tags, 73 albums, 1500 artists, 65000 tracks, and 18000 sound extracts (26 seconds each). This dataset was generate in the summer of 2005, and is used in the demo version of Ikoru[5] application.

In addition to the above, we have also collected data about music charts in Last.fm. The data contains song titles, bands, and their positions in the charts on weekly bases. This data is available from http://users.ecs.soton.ac.uk/mns2/charts/lastfm-charts-20050213-20070714.zip.

[4] http://www.zexe.net/SAOPAULO/

[5] http://demo.ikoru.net/

1.11 Armin Linke Photos

In a user study involving photos by the photographer Armin Linke, we collected data about 30 users, 4600 photos, 300 tags, and 4102 tag assignments. All stored in MySQL and also available in XML. The user study was carried out in May/June 2006. The photos are copyrighted, but the tags will be made available. This user study was part of an initial Ikoru testing.

1.12 Phenotypes/Limited Forms

Phenotypes/Limited Forms is an art installation that uses photos by the photographer Armin Linke and that has been on display at the Zentrum fur Kunst und Medien (ZKM[6]) in Karlsruhe, Germany, and at the Selective Knowledge[7] exhibition in Athens, Greece. We collected data about 8000 users, 1000 photos, 8000 tags, and 70000 tag assignments. The data gathering started in November 2007. The photos are copyrighted, but the tag assignments are available at `http://www.csl.sony.fr/~hanappe/phenotypes-20080425.txt`. The beta version of the Web interface for this project is located at `http://armin.hfg.ikoru.net/app/bookondemand/`.

1.13 Music Classifications

Over 40000 music files were annotated by experts with 800 boolean descriptors for every file. This dataset is currently being used by Sony to study methods for improving automatic music classification by combining audio analysis with statistical analysis of the descriptors; test machine learning on very large audio databases.

1.14 Cross-folksonomy users

We used Google Social API to search each of the 667K users in our del.icio.us dataset for their Flickr and Last.fm accounts, if there are any. This search produced 1998 users (users with del.icio.us, Flickr, and Last.fm accounts). We then collected all the tags and tagged resources for each of these users from each of their three accounts.

This set was used in the work on user profiling presented in (Szomszor et al., 2008a).

1.15 UK Music Singles Charts

UK music singles charts for the period 2002-2007 have been collected in RDF and can be downloaded from `http://users.ecs.soton.ac.uk/mns2/charts/top40-charts-20021208-20070715.zip`. This data is collected from Top40-charts[8]. This data consists of one file per week, where each file contains a list of song URIs that reached the top 40 in that particular week. Each song is linked to its title, band, a front cover image, and its position in the charts in that week.

[6] `http://www02.zkm.de/youser/index.php?option=com_content&task=view&id=80&Itemid=49`

[7] `http://www.itys.org/english/exhibitions/selective_knowledge.html`

[8] `http://www.top40-charts.com/`

Bibliography

Rabeeh Abbasi and Steffen Staab. Introducing triple play for improved resource retrieval in collaborative tagging systems. In *Proceedings of ECIR'08 Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2008)*, 3 2008. URL http://www.uni-koblenz.de/~abbasi/publications/Abbasi2008ITP.pdf.

Ivan Cantador, Martin Szomszor, Harith Alani, Miriam Fernandez, and Pablo Castells. Enriching ontological user profiles with tagging history for multi-domain recommendations. In *Proc. Workshop on Collective Semantics: Collective Intelligence and the Semantic Web (CISWeb 2008), in 5th ESWC, Tenerife, Spain*, 2008.

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences United States of America*, 104:1461, 2007. URL http://www.pnas.org/cgi/content/short/104/5/1461.

Klaas Dellschaft and Steffen Staab. An epistemic dynamic model for tagging systems. In *HYPERTEXT 2008, Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, 2008.

Olaf Görlitz, Sergej Sizov, and Steffen Staab. PINTS: Peer-to-peer infrastructure for tagging systems. In *Proceedings of the Seventh International Workshop on Peer-to-Peer Systems, IPTPS08*, Tampa Bay, USA, February 2008.

Miranda Grahl, Andreas Hotho, and Gerd Stumme. Conceptual clustering of social bookmarking sites. In *7th International Conference on Knowledge Management (I-KNOW '07)*, pages 356–364, Graz, Austria, SEP 2007. Know-Center.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference*, volume 4011 of *LNCS*, pages 411–426, Budva, Montenegro, June 2006. Springer. ISBN 3-540-34544-2. URL http://www.kde.cs.uni-kassel.de/hotho/pub/2006/seach2006hotho_eswc.pdf.

Martin Szomszor, Ciro Cattuto, Harith Alani, Kieron O'Hara, Andrea Baldassarri, Vittorio Loreto, and V. D. P. Servedio. Folksonomies, the semantic web, and movie recommendation. In *Bridging the Gap between Semantic Web and Web 2.0*, In Proceedings of 4th European Semantic Web Conference (in press), Innsbruck, Austria, 2007.

Martin Szomszor, Harith Alani, Ivan Cantador, Kieron O'Hara, and Nigel Shadbolt. Semantic modelling of user interests based on cross-folksonomy analysis. In *submitted to Int. Semantic Web Conf., Karlsruhe, Germany*, 2008a.

Martin Szomszor, Ivan Cantador, and Harith Alani. Correlating user profiles from multiple folksonomies. In *Proc. Int. Conf. Hypertext (HT08), Pittsburgh, PA, USA*, 2008b.