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)

D1.4 Data sets from folksonomic sites

Period covered: from 01/06/2008 to 31/08/2009
Start date of project: June 1st, 2006
Due date of deliverable: October 15th, 2009
Distribution: Public

Date of preparation: 15/09/2009
Duration: 39 months
Actual submission date: September 15th, 2009
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: PHYS-SAPIENZA
Lead contractor for these deliverables: TAGora partners

Contents

Chapter 1

Data collection from bibliographic reference sharing system

1.1 Introduction

The TAGora consortium fully understands the great benefit of publishing and sharing data in reusable formats. Making this data accessibly to the public is exceptionally valuable to all interested researchers and developers.

To this end, the project web site currently lists several datasets on a dedicated page[1], which contains links for downloading the various datasets, such as Delicious, Flickr, Last.fm, and various other datasets that TAGora used. Some of the data was anonymized to protect personal identities of users. The most valuable of the datasets are now available in RDF for download.

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 Delicious

Data from Delicious 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.

For legal concerns, this dataset is currently no available to the public. However, the scripts necessary for collecting data from this source are now available on the Data website. This will allow interested people to collect their data by giving their Delicious member name to this script, which in turn will collect information about their tags, tags resources, and frequency of tagging. This data will be returned in RDF.

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.

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

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.

1.4 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 data set currently consists of over 2.6 thousand users, 181 thousand bookmarks, 219 thousand publications, and over 816 thousand tag assignments. The dumps can easily be loaded into a MySQL database.

1.5 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[2] 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`.

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 Zexe.net: MOTOBOY

The dataset from the canal*MOTOBOY project[3], 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`.

[2] `http://demo.ikoru.net/`
[3] `http://www.zexe.net/SAOPAULO/`

1.8 Zexe.net: NoiseTube

To apply the zexe.net concept which is to find new tagging usages in the real world, we set up a new extension called NoiseTube for the last year. This extension is put in a new environmental context. It enables the general public to measure and annotate their exposure to noise pollution via their mobile phones. In the context of noise pollution, them measurement alone is not enough since we need to identify the causes of pollution to react on it. Since people are excellent at recognising noise sources, they can annotate the geolocated measurements regarding the cause or context via the mobile application before sending the enriched measurements to the platform. This kind of environmental tagging adds a semantic layer to the exposure map created by the participants.

1.9 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[4]) in Karlsruhe, Germany, and at the Selective Knowledge[5] exhibition in Athens, Greece, Arts Bienal, Sao Paulo, Brazil, Museum of Contemporary Art, Siegen, Germany. We collected data about 24 000 users, 2 400 photos, 17 000 tags, and 190 000 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.

1.10 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[6] 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[7] 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 in IMDB can be downloaded from http://users.ecs.soton.ac.uk/mns2/data/netflix_imdb_mapping.csv.

[4] http://www02.zkm.de/youser/index.php?option=com_content&task=view&id=80&Itemid=49
[5] http://www.itys.org/english/exhibitions/selective_knowledge.html
[6] http://www.imdb.com/interfaces
[7] http://www.netflixprize.com/download

1.11 Tag Senses

For several of our tag processing tools and services, we require information to help in tag sense matching and disambiguation. To this end, we processed a very large number of tags and DBpedia pages to generate the necessary knowledge for such tasks. Creating this dataset involved processing the whole XML dump of all Wikipedia pages and indexing all titles, then mining redirection and disambiguation links, then extracting term frequencies for each page. This dataset currently consist of over 51 thousand processed tags, over 197 thousand sense matchings, and nearly 6.5 million senses from DBpedia processing. This RDF dataset now contains well over 160 million triples.

Bibliography

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.