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)

D1.1 Data delivery from selected folksonomy sites and D1.3 Data delivery from selected recommender systems

---

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

Date of preparation: 31/05/2007
Duration: 36 months
Actual submission date: May 31st, 2007
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: “Sapienza” Università di Roma
Lead contractor for these deliverables: “Sapienza” Università di Roma and Univ. of Southampton

Chapter 1

Data collection from bibliographic reference sharing system

This document reports the summary information about the data collection and data delivery relatives to deliverables D1.1 and D1.3. The full information can be found at the wiki page of the TAGora project: `http://wiki.tagora-project.eu/DataRepository`.

1.1 BibSonomy benchmark dataset

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 Miranda Grahl (mgr@cs.uni-kassel.de) for access to our server on `https://www.kde.cs.uni-kassel.de/bibsonomy/dumps/2006-12-31.tar.gz`. Before starting the download, participants have to sign a license agreement in which terms of use are set up. The dataset includes data from approximately 400 users, 12.000(different)/140.000(all) tags and 39.000 resources and can easily be loaded into a MySQL database.

1.2 Del.icio.us benchmark dataset

Social bookmarking data have been crawled from del.icio.us and are available only for use within the TAGora project. The del.icio.us data have been crawled from November 10 till 24, 2006. The crawling was supported by all participants of the TAGora project and coordinated by the University of Kassel. The crawl was coordinated by a central server in Kassel. It monitored the 'recent posts page', resulting in an constantly updated list of user names. From this list, it distributed small chunks to over nearly 70 PCs, located all over the world, especially at Kassel, Rome, Koblenz, Southampton, Japan, Netherlands, Karlsruhe and Leipzig. These PCs crawled completely the corresponding user pages (including all follow up pages when a user page surpassed 5000 entries). Each PC waited a random time between 60 and 120 seconds before the next request, with an average delay of 90 sec. Globally, this resulted in an average of one request per second on the del.icio.us server. User pages that could not be downloaded completely were marked as 'incomplete' in the database.

Interested project members can contact Andreas Hotho (hotho@cs.uni-kassel.de) to get access to the dataset in html format via the server on `http://www.kde.cs.uni-kassel.de/crawldataset/`. The crawled data have also been made available in hdf-5 format by the University of Rome and are accessible on sismopil.phys.uniroma1.it. Interested people can contact Ciro Cattuto[1] to get

[1] ciro.cattuto@roma1.infn.it

access to the dataset.

Overall, the project members have internal access to 10GB compressed and 50GB uncompressed data to analyze and model evolutional behavior and structural information of social resource sharing systems. Overall, data from 667,128 users of the del.icio.us community with 18,782,132 resources, 2,454,546 tags (organized in 667 bundles) and 140,333,714 tag assignments were collected.

1.3 Last.fm dataset

Sony CSL crawled the Last.fm web site to obtain tagging information and related music extracts. The main objective to obtain this data was to examine the possibility to *ground* the tags through audio analysis. The crawl started from nine subscribers who actively tagged on the Last.fm web site. We retrieved their tags, the associated music data, and, finally, the audio extracts.

Last.fm is different from tagging sites for other media in that users not only tag music files, but also artists and albums. Because Sony was only interested in relations between tags and audio, we flattened the collected data: When an artist or album was tagged, the tag assignment was carried over to the music tracks of the artist or album, respectively. We retained the artist and/or album information as machine tags (see also Deliverable D2.2).

In total, the nine users provided 155 tags that mapped to 65,273 music tracks. However, for those tracks, only 18,444 audio extracts were available on the Last.fm site. This adds up to 18GB MP3 encoded audio (26 seconds per extract). The data can be obtained at `http://demo.ikoru.net`.

1.4 Flickr Benchmark Dataset

A crawler for the Flickr data was developed in Koblenz. The approach of the Flickr crawler is similar to the del.icio.us crawler but tailored to the peculiarities of Flickr. It is implemented in Java and uses an open source Java library to directly access the Flickr API. The backend for the crawler is a PostgreSQL database that stores the essential tagging information including the user-tag-photo relations and some additional information returned by the Flickr API.

Java was chosen as implementation language for the Flickr crawler because it can be run on any platform and allows for easy access to databases and the Flickr API via freely available libraries. The crawling strategy allows to run the crawl in parallel on distributed machines for predefined time intervals. The implementation is extensible to allow for easy adaption to new requirements for future crawling activities.

Extensive test with the Flickr API were conducted beforehand to ensure that all required data is correctly retrieved and a consistent dataset will be created. Some problems with the Flickr API were detected and workarounds implemented to avoid possible data irregularities. The actual crawl was then supervised by Koblenz and supported with the necessary infrastructure.

Helpful discussions with partners, especially with Kassel, concerning the crawling strategy and crawled data helped to improve the Flickr crawler. Due to the huge amount of data obtained from Flickr the crawl has been initially limited to Photos uploaded between January 2004 and January 2006. The crawling activity is still ongoing for this period but it can be expected that it will be finished June/July 2007. The latest snapshot of the ongoing crawling activity from mid of May contains 298,954 users, 24,599,875 photos, 1,553,253 tags and 110,345,103 tag assignments.

Depending on the data needs of the consortium it may be flexibly decided to extend the covered period of the data sets. For this purpose, it is possible to let the crawler retrieve only photos coming from a predefined period. Subsequent crawls benefit from previous crawling activities. For example, they can reuse the list of already known tags for retrieving further photos. The dataset is available at `http://kater.uni-koblenz.de/~klaasd/flickr/`.

1.4.1 Crawling Strategy

Several crawlers can be used in parallel. The crawling of the same photo by several crawler instances is avoided by assigning each crawler a certain period for which it is responsible. A period corresponds to approximately 1 week. If a crawler finishes a period it automatically gets a new, uncrawled period assigned. This way the crawling can be done in a unsupervised way.

When a new period is started, a random photo from that period is retrieved and stored in the database. This initialization of a period was especially important for the very first periods. After retrieving the first photo, the regular crawling activity starts. During crawling, all already discovered tags are stored in a central table of the database. A crawler issues search queries restricted to its period for each of those tags. This way, the crawlers benefit from each other because they also query those tags which were discovered by other crawlers in the meantime. A crawler proceeds with the next period if it reaches the end of the tag table.

1.5 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 from www.imdb.com/interfaces 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 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 17,770 movie titles in the Netflix dataset with their IMDB counterparts. An ontology, pictured below, for both the IMDB and Netflix dataset was conceived and then mapped to the relational database via the D2RQ mapping technology. The result is a large knowledge base on movies and movie ratings that supports semantic querying (for example through SPARQL).

As mentioned above, IMDB and Netflix data are available for download online so there is no need to write any screen scrapers. We will check if the downloadable files are updated every three months, and if so, we will regenerate the RDF and update the triple stores accordingly.

Interested partners can contact Harith Alani[2] or Martin Szomszor[3] to get access to the dataset.

[2] ha@ecs.soton.ac.uk
[3] mns2@ecs.soton.ac.uk