Supplementary Materials

Hierarchical Clustering-Based Outlier Detection
from Academic Social Tagging Data

Hang Dong[1,2], Wei Wang[2], and Frans Coenen[1]
[1]University of Liverpool,
{hangdong,Coenen}@liverpool.ac.uk
[2]Xi'an Jiaotong-Liverpool University,
Wei.Wang03@xjtlu.edu.cn

Material 1 The semantic assumptions and treatment for all special characters with real tag examples in Bibsonomy (referenced in "Handling special characters semantically" in Sect. 3 in the paper).

Material 2-3 The full cleaned dataset from Bibsonomy data after Step 4 (referenced in Table 1 in the paper) containing 2502 multiword tag groups (Material 2) and 14,877 single tag groups (Material 3).

Material 4 The reduced dataset after Step 5 (referenced in Table 1 in the paper) by selecting only tag groups annotated to publications, containing 15,647 multiword and single tag groups.

Material 5 (svd-2000) The SVD-reduced matrix (15,647*2000) created from Material 4, using the binary resource-based representation method, where each row vector corresponds to a tag group and each entry in the row vector corresponds to whether tags in the tag group were used to annotate a resource (if yes then 1, no then 0). The dimension of the matrix is reduced from 301,669 (number of distinct resources after Step 5 in Table 1 in the paper) to 2000 using SVD, with around 80% of eigencomponents retained. All row vectors in Material 5 correspond to the tag groups of the same rows in Material 4.

Material 6 The clustering results and candidate outliers with data from Material 4 represented using a binary resource-based matrix reduced to 2000 dimensions after Singular Value Decomposition.

Material 7 The candidate outliers per clustering technique (Sheet 1 from Columns A to O), distinct candidate outliers from all clustering techniques (Sheet 1 Column P), human evaluation results (Sheet 1 Columns Q to W); and *precision*, *recall*, *F-measure* for each clustering technique (Sheets 2 to 4).

Notes:

1. The original raw Bibsonomy Dataset can be requested from http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/. For this research, the file "2015-07-01.tgz" (227MB) is used as the input of the Data Cleaning workflow.

2. In Material 2-4, each line of text represents a tag group, having the form below.

```
[language] Standard_Tag_Form: TagA TagB TagC … TagN MetricGroups
isReliable:false confidence:confidence_percentage
```

, where **`MetricGroups`** include 6 metrics for single tag groups (Material 3)

```
Tag_Frequency Number_of_Distinct_Resources Number_of_Distinct_Users
Inverse_Resource_Frequency User_Frequency_Inverse_Resource_Frequency
Normalised_Annotation_Frequency
```

, and another metric called `Multiword_Likelihood` for multiword tag groups (Material 4). So multiword tag groups have 7 metrics in **`MetricGroups`**.

```
Tag_Frequency Number_of_Distinct_Resources Number_of_Distinct_Users
Inverse_Resource_Frequency User_Frequency_Inverse_Resource_Frequency
Normalised_Annotation_Frequency Multiword_Likelihood
```

The language in [] for each tag group is obtained using Google Translation on April 2016.

Below is the explanation of each metric with its lower-letter abbreviation (used as head column in Material 6).

- Tag_Frequency (Nt): The number of annotation of tags in a tag group, for any users and resources.

- Number_of_Distinct_Resources (Nr): The distinct number of resources annotated using any of all the tags (including standard tag form) in a tag group.

- Number_of_Distinct_Users (Nu): The distinct number of users who annotated any of all the tags (including standard tag form) in a tag group.

- Inverse_Resource_Frequency (irf): Similar to Inverse Document Frequency for information retrieval, irf was designed to measure the information a tag group provides, calculated as the logarithmically scaled inverse fraction of the documents that contain the tag, irf = log10(N/Nr), where N is the total number of distinct resources in the cleaned dataset. **This metric is not used in the Data Cleaning workflow**.

- User_Frequency_Inverse_Resource_Frequency (ufirf): Similar to Term Frequency-Inverse Document Frequency for information retrieval, ufirf was designed to measure the importance of a tag group. ufirf is calculated as the product of Nu and irf, ufirf = Nu*irf. **This metric is not used in the Data Cleaning workflow** since it is not better than simply using Nu as a measure.

- Normalised_Annotation_Frequency (Norm-af): The rate of Tag Frequency of a tag group to the Number of Distinct Resources of the same tag group,

  Norm-af = Nt/ Nr.

  This metric is useful to measure the significance of a tag group according to the reputation of resources: the extent of users annotating a same resource using tags in a

tag group. The metric is therefore contextual, i.e. only works when the tag groups that are popular in the dataset. When the Nu of a tag group reach a certain threshold (Nu >= 3), the higher the Norm-af, the more significant the tag group.

- Multiword_Likelihood (mwl): Some tags like "database", "radioactive", "multilingual" contain two lexemes rather than two words; we call them *multi-lexeme single tags*. The standard tag form for these expressions should be only letters, rather than with a hyphen inside such as "data-base". It is necessary to make distinction of the two types to generate a more accurate standard tag form. To distinguish *multi-lexeme single tags* like "database" to other proper multiword tags like "data mining", we propose a metric called Multiword Likelihood.

  mwl = sum of frequency (number of annotations) of *explicit multiword tags* in a multiword tag group / sum of frequency of all tags in a multiword tag group,

  where an *explicit multiword tag* is either (1) a tag that have an underscore between two letters or numbers; or (2) a tag that have the pattern of xXx showing one capital letter between 2 lowercase letter.

  If mwl for a tag group is below a threshold, then we assume that the standard tag of this multiword tag group includes only one word but more than one lexeme. The idea behind this metric is inter-subjectivity or users' collective intelligence. In this way, we can precisely determine the tag "database" rather than "data_base" as the standard tag form.

  For example, the multiword tag group labelled by frequency in each tag form.

  "Time_Management: timemanagement(54) TimeManagement(26) time_management(9) time-management(5) Time_management(4) time.management(1)", mwl = (26+9+5+4)/(54+26+9+5+4+1) = 0.44.

  However, for the other tag group

  "Data_Set: dataset datasets Dataset data-set DataSet Datasets data-sets DataSets datset data_set DATASET Data_Sets", the mwl is only 0.04.

  This shows that Date_Set is more likely to be a multi-lexeme single word, but Time_Management is more likely be a multiword. Therefore the standard tag form for Data_Set should actually be "dataset", without an underscore, while the Time_Management is unchanged.

3. In Material 6, the actual number of tag groups that were clustered is 15,402 out of 15,647 tag groups after Step 5 in Table 1 in the paper. This is because 245 tag-group vectors having length smaller than 0.0001 were filtered out. When the vector length is too small, it is not possible to achieve a proper Cosine Similarity. The sheet 1 in Material 6 contains 15,402 tag groups, and the original 15,647 tag groups are in the sheet 2 and Material 4.

4. In Material 7, the results of final1 (Sheet 1 Column V), achieved by selecting final outliers as those marked at least 3 times as such out of 5 participants, were presented in the paper. The set final2 (Sheet 1 Column W), outliers marked at least 4 times, was also used for calculating *precision*, *recall* and *F-measure*, as shown in the Sheet 4.