Searching and Browsing Tag Spaces Using the Semantic Tag Clustering Search Framework

Erasmus University Rotterdam

Introduction

Many Web services enable users to label content on the Web by means of **tags**. A well-known application that makes use of tags is **Flickr**. Users that are registered on the Flickr Web site can upload photographs and assign tags to them. As the user has no restrictions on the tags that can be used, tags are prone to errors and ambiguity, and thus have their **limitations**:

- **Typographical errors** and **syntactic variations**, i.e., different tags having the same meaning (e.g., *waterfall*, *waterfal, water-fall*, etc.);
- **Synonyms**, i.e., semantic relatedness *(*e.g., *interior*, *inside, indoor*, etc.)*;*
- **Homonyms**, i.e., tags that have multiple meanings (e.g., a*pple, orange, mouse,* etc.).

Framework

As a solution to the previously introduced problems related to tagging, we define the **Semantic Tag Clustering Search** (STCS) framework, which consists of **two parts**:

- The first part of the framework deals with **syntactic variations**;
- The second part of the framework is concerned with deriving **semantic clusters**.

For **syntactic variation detection**, we employ the **normalized Levenshtein distance**, i.e., the minimum number of edits needed to transform one string into the other, divided by the maximum string length. For **clustering**, an undirected graph is used, which contains weighted edges. Weights are based on a **weighted average** of the **normalized Levenshtein distance** (more suitable for long tags) and the **cosine similarity** (representative for short tags) between two tags.

Semantic clustering is based on **semantic relatedness**. To measure the semantic relatedness between tags, we use the **cosine similarity** based on co-occurrence vectors. We consider non-hierarchical clusters, where we select the method proposed by **Specia and Motta (2007)**, which has **two steps**:

- Merge clusters if one cluster contains the other cluster;
- Otherwise, merge clusters if they differ within a small constant margin with respect to the size of the smaller cluster (static threshold).

A **drawback** is the **sensitivity** to the **size** of the smaller cluster. Hence, we **adjust** this method:

- Merge clusters if one cluster contains the other cluster;
- Otherwise, use in a disjunction:
    - Merge clusters when the average cosine is above a threshold (semantic relatedness);
    - Merge clusters if they differ within a small margin that depends on the square root of the size of the smaller cluster (dynamic threshold).

Performance

On a test set on Flickr data with 200 randomly chosen tag combinations, we identify an **error rate** of **5%** for **syntactic variation** detection. For **semantic clustering** evaluation, we select 100 random clusters (which have 458 tags in total), and obtain an **error rate** of **9.6%** (44 misplaced tags). We improve the original algorithm, with a semantic error rate of 13.1%, by 26.7%. Also, we identify 75.5% more clusters (739 against 421). The algorithm identifies clusters that contain tags that are translations of concepts in different languages (e.g., *springtime*, *primavera*).

Conclusions

We have proposed the **Semantic Tag Clustering Search** (STCS) framework for building and utilizing semantic clusters based on information retrieved from tags. We **remove syntactic variations** and **create semantic clusters**. We obtain a syntactic error rate of 5%, and, compared to an existing approach, a better semantic clustering error rate of 9.6%.

Jan-Willem van Dam
jwvdam@gmail.com
Damir Vandic
damir3004@gmail.com
Frederik Hogenboom
fhogenboom@ese.eur.nl
Flavius Frasincar
frasincar@ese.eur.nl

Econometric Institute
Erasmus School of Economics
Erasmus University Rotterdam
P.O. Box 1738, NL-3000 DR
Rotterdam, the Netherlands
Phone: +31 (0)10 408 8907
Fax: +31 (0)10 408 9031
http://www.eur.nl/english/

ERASMUS UNIVERSITEIT ROTTERDAM