Collaborative Tagging for Word Relationships Mining

TOMÁŠ MICHÁLEK*

Slovak University of Technology

Faculty of Informatics and Information Technologies

Ilkovičova 3, 842 16 Bratislava, Slovakia

michalek07@student.fiit.stuba.sk

As amount of data on the internet is growing faster and faster, information and data are becoming hard to discover and explore. It is nearly impossible or possible in a very small scale to process unstructured text, photo or video content. Nowadays searching is heavily based on looking up a key words. It's like playing darts and hoping to hit the right words. Even then we won't get the information we are looking for, instead we get an amount of links referring to resources (articles, pages, etc.) where the desired information is wrapped up with a lot of data, which are at the moment useless for us. So even answering simple question or looking up for single fact can be a very long process.

We can split this content into several basic categories by it's format. Structured or mostly structured text, which can be relatively easily processed automatically in a very large scale like RSS feeds.

Second category are resources with unstructured text data. This resources often contains also a lot of text vapid to user which can distort a search results and become very misleading. Delisious.com is online bookmarking web service. Users are using this system to bookmark interesting content they found on internet or to discover another based on their interests. Users can add to every bookmarked link tags to make it more discoverable and searchable. As many people mark resources we can extract more information telling us something about relations between this marks and resources.

Third category is content that is very hard to process automatically, like photo and video content. Pages like youtube.com or Flickr.com have to deal with this problem. Without people marking this content we won't be able to search in it.

In my bachelor project I am focusing on these systems, describing methods of tag relations mining and behaviour and nature of these systems. Despite of the content we are tagging all these systems have common structure creating a tripartite graphs. One

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2010 PeWe Workshop, April 18, 2010, pp. 63–64.

tagging instance is an edge connecting user, resource and tag. Another common feature of these systems is that tag distribution between resources follows a power law.

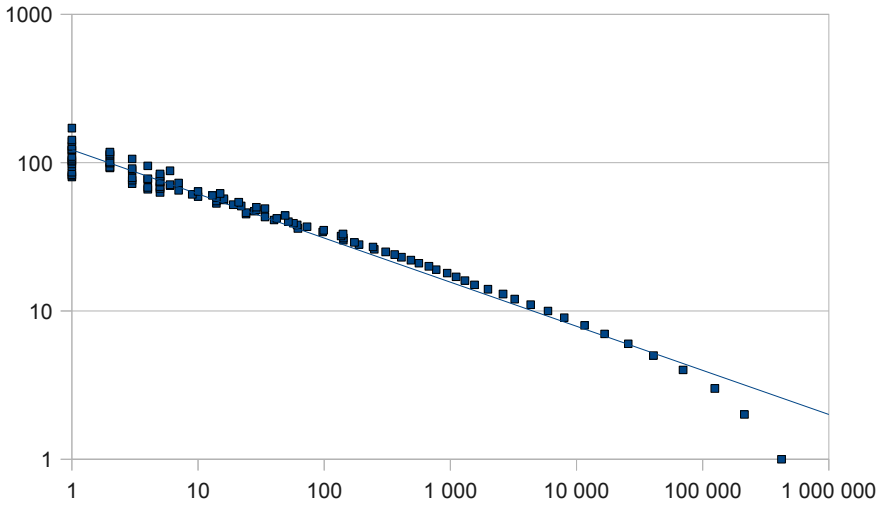

Figure 1. Tag distribution on resources. Dataset from delicious.com

On Figure 1. is displayed tag distribution of tags. On the x ax is number of resources; on y is marked how many tags it contains. Both axes are in logarithmic scale. This means that 90% of resources have between one and five tags. If we are able to discover relations between tags, we can extend these resources by new words making content more accessible to users.

But there can also be another motivation for word relationship mining. If we know a relations between words we can better understand a content, meaning that we can process unstructured text nearly same way as structured. This allow to extract from text not just key words but also facts and information. This makes a huge difference in creating a search query. Instead of trying to hit a key words and scrolling through a great amount of links we can ask a question same way we are asking another person in real life and just get an short and factual answer.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

[1] Halpin, H., Robu, V., Shepherd, H. The Complex Dynamics of Collaborative Tagging. WWW 2007 / Track: E*-Applications