Frequency- and Content-Based Tag Cloud Font Distribution Algorithm

Ágnes Bogárdi-Mészöly, Takeshi Hashimoto, Shohei Yokoyama, Hiroshi Ishikawa

Abstract—The spread of Web 2.0 has caused user-generated content explosion. Users can tag resources to describe and organize them. Tag clouds provide rough impression of relative importance of each tag within overall cloud in order to facilitate browsing among numerous tags and resources. The goal of our paper is to enrich visualization of tag clouds. A font distribution algorithm has been proposed to calculate a novel metric based on frequency and content, and to classify among classes from this metric based on power law distribution and percentages. The suggested algorithm has been validated and verified on the tag cloud of a real-world thesis portal.

Keywords—Tag cloud, font distribution algorithm, frequency-based, content-based, power law.

I. INTRODUCTION

WITH the appearance of Web 2.0 and the spread of social media sites users became from passive spectators to active content generators. Users can interact and collaborate with each other in virtual communities. Nowadays lots of social sites exist for various purposes: collaborative projects (Wikipedia), blogs (Twitter), content communities (YouTube), social networking sites (Facebook), virtual game worlds (World of Warcraft), virtual social worlds (Second Life), etc. There are numerous books dealing with this topic denoted for introduction and marketing, for research, etc. Various significant companies have research groups for social computing: Microsoft, IBM, HP, etc.

Usually on these sites users can assign tags to resources to describe and organize them. This tagging process establishes associations between tags and resources, which can be applied to navigate to resources by tags, as well as, to tags based on related tags, etc. [1]. With tagging a folksonomy (folk (people) + taxis (classification) + nomos (management)) evolves, which is the vocabulary of tags emerged by the community [2]. The classification scheme is non-hierarchical, plain. The vocabulary is open, predefined tags do not exist, new words can be chosen freely as tags. Thus, its size tends to grow indefinitely. Moreover, it is incomplete and inconsistent. Thus, in connection with social tagging several challenges have emerged [3].

The paper is organized as follows. Section II covers the background. Section III introduces the experimental environment and open issues. Section IV proposes a novel

Á. Bogárdi-Mészöly is with the Department of Automation and Applied Informatics, Budapest University of Technology and Economics. (e-mail:agi aut.bme.hu).
T. Hashimoto is with the Department of Electrical and Electronics Engineering, Shizuoka University.
S. Yokoyama is with the Department of Computer Science, Shizuoka University.
H. Ishikawa is with the Department of Information and Communication Systems, Tokyo Metropolitan University.

font distribution algorithm. Section V shows the experimental results. Finally, last section reports the conclusions and future work.

II. BACKGROUND

In this section, definitions related to tagging are discussed. Tags are user-defined informal and personal strings, short descriptions related to resources, keywords associated with resources. They are helpful in browsing and searching. Resources are such identities which can be tagged, such as text, image, audio, video, document, etc. Tagging is the process of assigning existing and new tags to resources. Tag recommendation systems exist to help users in tagging based on own tags or tags of other community members.

Tagging can be simple or collaborative. In case of simple tagging users assign tags only to their own resources. In case of collaborative tagging many users tag the same resource, each user can tag it with her/his own tags in her/his own vocabulary. In our tag clouds, simple tagging is applied.

There are lots of different kinds of tags: content-based, context-based, attribute, ownership, subjective, organizational, purpose, factual, personal, self-referential, tag bundles, etc. Furthermore, users have various motivations for tagging: future retrieval, contribution and sharing, attract attention, play and competition, self-presentation (self-referential tags), opinion expression, task organization, social signaling, money, technology ease, etc. [3]. In our tag clouds, tags are content-based, tagging is motivated by contribution and sharing. With content-based tags the actual content of the resources can be identified. By the contribution and sharing as motivation, tags describe resources, and add them to conceptual clusters or refined categories for known and unknown audience.

Tag clouds are visually depicted tags in order to facilitate browsing among numerous tags and resources. It gives rough impression of relative importance of each tag within the overall cloud. In this paper such tag clouds are investigated. In some situations to answer various questions browsing in tag clouds are more useful than searching [4]. Search interface is preferred if the needed information is specific. Tag clouds are preferred if the sought information is more general.

For this reason, the visualization of tag clouds is one of the most important and complicated consideration [5]. Tag clouds have two dimensional representations. Tags can be ordered alphabetically, randomly, based on semantic similarity or any kind of clusters [6], [7], [8]. Relevant tags can be visually emphasized using visual properties such as shape, color,

position, etc. Each tag cloud is visualized in its own unique way. The basis of the used methods is similar, but there are no two tag clouds whose visualization is the same. Numerous font distribution algorithms exist [9], [10]. In our tag clouds all tags are represented simply ordered by alphabetically and visually weighted by letter size.

In social networks power laws occur many times in many contexts [3]. A random variable is distributed according to a power law when its probability density function is given by $x^{-\gamma}$, where $\gamma > 1$ is a constant parameter called exponent or scaling parameter, typically in the range of $2 < \gamma < 3$, and $x \geq x_{min}$ [11].

In our previous work, complete tag and topic recommendation systems have been established [12]. The tag recommendation system has three main steps. In the first step the vocabulary has been refined, namely, certain spelling and clerical errors have been corrected, in addition, certain tags have been contracted. In the second step the reference counts have been enhanced. In the third step the font distribution algorithm has been improved. The topic recommendation system has three main steps. In the first step a special graph has been constructed from tags. In the second step the reference count of each node has been evaluated in order to identify topics. In the third step the improved font distribution algorithm of the proposed tag recommendation system has been applied and visualization has been introduced. In this paper, the second and third steps of the tag recommendation system have been improved.

III. Experimental Environment

In this section, the experimental environment and the emerged open issues are introduced.

The Faculty of Electrical Engineering and Informatics of the Budapest University of Technology and Economics has a thesis portal to maintain the whole workflow for all theses of the faculty starting from creation of new thesis ending with review upload [13].

This portal is implemented as a three-tier ASP.NET web site. The presentation layer is in HTML and jQuery. In the business logic layer there are C# classes. In the data access layer LINQ and stored procedures are used mixed. The database is in Microsoft SQL Server. For fast visualization of tag clouds, the tags are stored in a cache.

The provided algorithm has been implemented in SQL stored procedures, C# classes using LINQ, and MATLAB functions.

On existing tag clouds it has to be executed once in a maintenance phase. After that or for newly created tag clouds it can be applied in two different ways: the whole algorithm is executed periodically only in maintenance phases, or only the related part is executed when a tag/abstract is created/modified/deleted.

Each thesis includes an uploaded document, an abstract which is a short extraction of the completed work, some keywords in other words tags, etc. Tags are assigned by authors to describe and organize them in order to be helpful in browsing and searching. The portal has tag clouds in English and Hungarian languages.

Fig. 1   Original tag cloud of most popular tags

The originally applied font distribution algorithm uses weights based on number of classes, average, maximum, minimum reference counts, and tags are classified to four classes.

Only some specific keywords are assigned to a thesis. Firstly it follows from "specific" that general keywords are less emphasized, since there are a lot of less emphasized, different but in similar topic, longer, specific keywords. For example, see the following tags: mobile, mobile application, mobile application development, mobile communication, mobile network, mobile phone, mobile robot, mobile software development, etc. All mentioned tags contain the word "mobile". Thus, the emphasis of tag "mobile" can be improved considering all mentioned tags. Secondly it follows from "some" that not all topics of a thesis are covered with its keywords, but more can be found in its abstract and uploaded document.

Furthermore, with the originally applied font distribution algorithm too many tags are emphasized in case of smaller tag clouds like most popular tags, tag cloud of a supervisor, etc. (see in Fig. 1), In addition, too few tags are emphasized in case of bigger tag clouds like all tags, tag cloud of a department, etc. (see in Fig. 2).

Our goal is to handle the above mentioned issues by frequency- and content-based metric calculation as well as power law distribution- and percentage-based classification among classes in order to improve tag cloud visualization.

IV. Improved Font Distribution Algorithm

This section proposes an improved font distribution algorithm in order to improve tag cloud visualization. Table I summarizes the notations of equations and algorithms. Firstly,

microchip · microcontroller · microcontroller programming · microfluidics · microprocessor · Microsoft · Microsoft .NET · Microsoft .Net 4.0 · Microsoft .NET 4.5 · Microsoft SharePoint · Microsoft SQL Server · microwave circuit design · middleware · migration · mikrocontroller · mikrokontroller · MIMO · Mininet · miniwave · MIPv6 · MMORPG · mobil · mobile · mobile application · mobile application development · Mobile communication · mobile network · mobile networks · mobile phone · mobile robot · mobile software development · mobility · mock object · Modbus · model · model based testing · model checking · model driven software development · Model-Based Design · model-driven development · modeling · modelling · modular · module · MongoDB · monitoring · Monte Carlo · motion tracking · motor control · MP3 · MPEG · MPEG-2 transport stream · MPPT · MSP430 · MSZ 274 · MSZ EN 62305 · multiagent learning · multiagent system · multiagent system control · multiagent systems · multicast · multihop · multilateration · multimedia · multiplatform · multiplayer · Multithreading · multi-touch · music · music player · MVC · MVVM · MySQL · National Instruments · natural language processing · navigation · Nd:YAG laser · NEMO · netflow · network · Network Attacks · network coding · network design · network development · Network management · Network Mobility · network monitoring · network planning · network security · network simulation · network topology · neural network · NFC · node.js · noise · non-linear system · non-photorealistic rendering · NoSQL · NOX · NXT · Objective-C · observer · OFDM · OFDMA · offline · off-line signature verification · Ogre 3D · Ogre3D · OLAP · OMNeT · Omnet++ · online · ontology · open source · Open vSwitch · OpenCL · OpenCV · OpenFlow · OpenGL · OpenGL ES · OpenGL ES 2.0 · OpenNebula · OpenWRT · optical · optical flow · optical network · optical networks · optimalization · optimization · Oracle · oscilloscope · OSGi · OSPF · overvoltage · OWL · P2P · P300 · parallel programming · parameter estimation · Partial Volume Effect · path planning · PCB · PCB design · PCS7 · peer-to-peer · performance · performance analysis · performance optimization · permanent magnet synchronous · PET · Petri net · Petri-net · photovoltaic cell · PHP · PHP programming ·

Fig. 2 Part of original tag cloud of top 1000 tags

TABLE I
NOTATIONS OF EQUATIONS AND ALGORITHMS

Notation	Description
τ	set of tags
π	set of papers
t_i	i^{th} tag
p_i	i^{th} paper
a_i	abstract of ith paper
$t_i \to p_j$	tag t_i is assigned to paper p_j
rc_i	reference count of a tag t_i
sc_i	summed reference count of a tag t_i
ac_i	abstract reference count of a tag t_i
fc_i	frequency- and content-based metric of a tag t_i
sc_{max}	maximum value among all summed reference counts
ac_{max}	maximum value among all abstract reference counts
w_{sc}	weight related to summed reference count
w_{ac}	weight related to abstract reference count
co_i	original class of ith tag
ci_i	improved class of ith tag

a novel metric has been established combining frequency- and content-based approaches. Secondly, using this metric a classification among classes has been provided based on power law distribution and percentages.

The reference count of a given tag is defined as (1). This count has been enhanced to the sum of reference counts of all tags which contain the given tag (see the motivation in the previous section related to phrase "specific" and tag "mobile"). This summed reference count can be seen in (2).

$$rc_i = |\{\forall p_j \in \pi, \text{ where } \exists\ t_i \to p_j\}| \qquad (1)$$

$$sc_i = \sum_{\forall t_j \in \tau_s} rc_j,$$
$$\text{where } \tau_s = \{\forall t_j \in \tau, \text{ where } t_j \text{ contains } t_i\} \qquad (2)$$

Moreover, a new content-based count has been introduced, which is the sum of papers whose abstracts contain the given tag (see the motivation in the previous section related to phrase "some"). This abstract reference count can be seen in (3).

$$ac_i = |\{\forall p_j \in \pi_a\}|,$$
$$\text{where } \pi_a = \{\forall p_j \in \pi, \text{ where } a_j \text{ contains } t_i\} \qquad (3)$$

The summed and abstract reference counts have been scaled with their maximum values, and can be weighted arbitrary. This frequency- and content-based metric is shown in (4).

$$fc_i = w_{sc} \cdot \frac{sc_i}{sc_{max}} + w_{ac} \cdot \frac{ac_i}{ac_{max}},$$
$$\text{where } w_{sc} + w_{ac} = 1 \qquad (4)$$

This metric seems to obey a power law. The investigation has been performed with MATLAB functions provided by [14]. Firstly, a power law distribution has been fitted to it. In addition, its scaling exponent and lower bound of power law behavior parameters have been estimated by maximum likelihood. Furthermore, the uncertainties in the estimated parameters have been estimated, as well. Then, the test whether the power law is a plausible fit to the empirical data set has been performed graphically and numerically, as well. The graphical test has been shown that the empirical data follows a straight line but deflects from it at the end, this means that the empirical data has a longer tail than the estimated theoretical power law distribution. The numeric test has been performed with the help of a hypothesis test using Kolmogorov-Smirnov statistic. Therefore, the power law is a plausible hypothesis for the frequency- and content-based metric.

Since the frequency- and content-based metric obey a power law, a power law distribution- and percentage-based approach has been led to correct visual impression. The proposed distribution algorithm divides the area of the given power law function (5) according to arbitrary percentages.

$$\frac{\max(fc)^{-\gamma+1} - \min(fc)^{-\gamma+1}}{-\gamma+1} \qquad (5)$$

V. EXPERIMENTAL RESULTS

The proposed algorithm (Section IV) has been validated and verified on a tag cloud of a real-world thesis portal (Section III). This section presents the experimental results.

In Table II the arbitrary chosen parameters and calculated values can be seen. With the chosen weight of the frequency- and content-based metric the summed reference count is emphasized considering less the abstract reference count. The distribution of the frequency- and content-based metric can be

TABLE II
ARBITRARY CHOSEN PARAMETERS AND CALCULATED VALUES

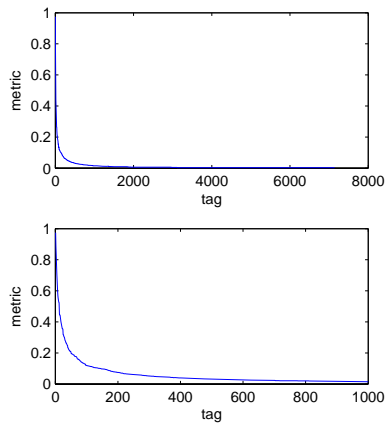|	
w_{sc}	4/5
w_{ac}	1/5
γ	2.1036
x_{min}	0.019
number of classes	4
percentage for class 4	0.5%
percentage for class 3	1.5%
percentage for class 2	3%

Fig. 3   Distribution of frequency- and content-based metric

observed in Fig. 3. The scaling exponent and lower bound have been calculated during demonstrating the power law behavior of the frequency- and content-based metric. The number of classes and the related percentages for distributing among classes have been chosen according to practical experience.

In Table III the importance of the proposed summed and abstract reference counts can be observed in numbers, which shows their considerable effect. The calculated values of top 25 tags can be seen in Table IV. With the improved algorithm there are no decrease in class belonging, only some originally important tags and newly recognized as frequently used important tags are emphasized more.

Table V displays the distribution among classes using the original and improved algorithms in case of top 1000 and 100 tags. The original algorithm classifies regardless of the tag cloud size. Recall that too many tags are emphasized in case of smaller tag clouds, too few tags are emphasized in case of bigger tag clouds. The improved algorithm ensures better classification among classes considering the tag cloud size.

In Figs. 4 and 5 the resulted tag clouds of most popular and top 1000 tags are depicted. For comparison the original tag clouds are illustrated in Figs. 1 and 2. In the original one

TABLE III
CHANGE IN COUNTS

	Number of tags
Total	7140
$sc > rc$	1113
$ac > 0$	4424

TABLE IV
TOP 25 TAGS

tag	rc	sc	ac	fc	co	ci
system	2	271	1540	0.974	1	4
Android	197	280	200	0.826	4	4
network	20	260	587	0.819	1	4
control	44	247	655	0.791	2	4
application	12	201	970	0.700	1	4
development	8	167	1204	0.634	1	3
data	3	165	1103	0.615	1	3
Java	87	193	158	0.572	2	3
design	8	146	1064	0.555	1	3
software	12	142	1015	0.538	1	3
mobile	30	162	484	0.526	1	3
web	47	167	364	0.524	2	3
management	5	142	331	0.449	1	3
processing	1	140	336	0.444	1	3
power	1	135	445	0.444	1	3
.NET	106	145	89	0.426	3	3
measurement	15	125	429	0.413	1	3
microcontroller	90	134	139	0.401	2	3
robot	41	134	96	0.395	2	3
test	9	103	673	0.382	1	3
Windows	7	121	135	0.363	1	2
simulation	75	111	303	0.356	2	2
game	15	118	146	0.356	1	2
energy	3	111	279	0.353	1	2
framework	27	95	358	0.318	1	2

TABLE V
CLASS BELONGING IN CASE OF TOP 1000 AND 100 TAGS

	1000 tags		100 tags	
	original	improved	original	improved
in class 4	1	5	1	1
in class 3	1	15	1	2
in class 2	29	30	29	3
in class 1	969	950	69	94

in case of smaller tag cloud too many tags are emphasized, in addition, in case of bigger tag cloud almost all tags are in unite. However, the improved tag cloud is much more perspicuous, the popular tags can be identified easily.

VI. CONCLUSION AND FUTURE WORK

The visualization of huge tag clouds is one of the most important and complicated consideration. In our thesis portal tag clouds have a very important role to facilitate browsing and searching among numerous tags and theses.

Fig. 4   Improved tag cloud of most popular tags

Fig. 5   Part of improved tag cloud of top 1000 tags

In this paper a novel font distribution algorithm has been proposed to improve tag cloud visualization, namely, a novel frequency- and content-based metric has been established, in addition, a power law distribution- and percentage-based classification among classes has been presented.

The proposed algorithm has been implemented in SQL stored procedures, C# classes using LINQ, and MATLAB functions. Its validation and verification have been performed on the tag cloud of a real-world thesis portal.

The further improvement of visualization is a subject of future work by reorganizing tag clouds combining co-occurrence- and content-based approaches.

ACKNOWLEDGMENT

REFERENCES

[1] S.A. Golder, B.A. Huberman, *The Structure of Collaborative Tagging Systems*, 2005.
[2] D.H. Pink, *Folksonomy*, New York Times, December 11, 2005.
[3] C.C. Aggarwal, *Social Network Data Analytics*, 1st Edition, Springer, ISBN 978-1-4419-8461-6, 2011.
[4] J. Sinclair, M. Cardew-Hall, *The folksonomy tag cloud: when is it useful?*, Journal of Information Science, Vol. 34(1), pp. 15–29, 2008.
[5] S. Lohmann, J. Ziegler, L. Tetzlaff, *Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration*, INTERACT, Part I, LNCS 5726, pp. 392–404, 2009.
[6] Y. Hassan-Montero, V. Herrero-Solana, *Improving Tag-Clouds as Visual Information Retrieval Interfaces*, International Conference on Multidisciplinary Information Sciences and Technologies, Merida, Spain, 2006.
[7] J. Salonen, *Self-Organizing Map Based Tag Clouds*, OPAALS Conference, 2007.
[8] A.P. García-Plaza, A. Zubiaga, V. Fresno, R. Martínez, *Social Media Mining and Social Network Analysis: Emerging Research*, Chapter 9, Tag Cloud Reorganization: Finding Groups of Related Tags on Delicious, pp. 140–155, 2013.
[9] kentbye's blog, *Tag Cloud Font Distribution Algorithm*, http://www.echochamberproject.com/node/247, June 24, 2005.
[10] K. Hoffman, *In Search Of ... The Perfect Tag Cloud*, http://files.blog-city.com/files/J05/88284/b/insearchofperfecttagcloud.pdf.
[11] A. Clauset, C.R. Shalizi, M.E.J. Newman, *Power-law distributions in empirical data*, SIAM Review, Vol. 51(4), pp. 661–703, 2009.
[12] Á. Bogárdi-Mészöly, A. Rövid, H. Ishikawa, S. Yokoyama, Z. Vámossy, *Tag and Topic Recommendation Systems*, Acta Polytechnica Hungarica, Vol. 10, No. 6, pp. 171–191, 2013.
[13] Thesis Portal, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, http://diplomaterv.vik.bme.hu/en/.
[14] A. Clauset, C.R. Shalizi, M.E.J. Newman, *Implementation for Article of Power-Law Distributions in Empirical Data*, http://tuvalu.santafe.edu/~aaronc/powerlaws/