Engineering Applications of Artificial Intelligence 64 (2017) 378–390

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Fast algorithms for online construction of web tag clouds

Jakub Marszałkowski, Dariusz Mokwa, Maciej Drozdowski *, Łukasz Rusiecki, Hubert Narożny

Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland

ARTICLE INFO

Keywords:
Web engineering
Tag clouds
2D packing
Data visualization
Metaheuristics
Automatic web page generation

ABSTRACT

In this paper tag cloud construction for web exposition is studied. Construction of a tag cloud must simultaneously solve at least three interdisciplinary engineering problems: modeling and controlling graphics aesthetics, solving discrete two-dimensional layout optimization problem, and all these must be done on computationally constrained browser platform. We analyze the design choices in the earlier tag cloud studies and provide a taxonomy of algorithmic approaches to tag cloud building. Then, the design requirements for tag clouds on websites are defined. We propose to quantify tag cloud aesthetics by use of a novel objective function based on the rules of typography. Tag cloud construction is formalized as a combinatorial optimization problem with an irregular objective function. A set of algorithms is proposed and evaluated on a collection of tag sets from popular web pages. The methods that meet constraints of the browser platform are chosen.

1. Introduction

Running web systems and developing web applications are new branches of industry offering a host of engineering and research challenges. These cover, e.g., classical performance tuning problems (Marszałkowski et al., 2011), novel e-business applications (Lopez-Loces et al., 2016), website layout optimization for good structure and advertisement fit (Marszałkowski and Drozdowski, 2013), content analysis and fast delivery (Kudelka et al., 2014; Marszałkowski et al., 2016), techniques for content interpretation and exploitation (Spyrou and Mylonas, 2016).

In this paper we analyze the problem of constructing visually acceptable tag clouds for web pages. Basically, *tags* are phrases representing textually some set of objects. Tags can be, e.g., words and phrases summarizing content of a web page or a photograph, labels for best-sellers, keywords in news, social media or scientific publications. Each tag has certain importance which is expressed in relation to other tags. Typically, tag importance is given as a number. A *tag cloud* is a graphical depicting of the tags projected onto a plane. A key requirement is that tags with high importance should be prominently visible in the tag cloud. Commonly, important tags are simply bigger. An example tag cloud from Flickr website is shown in Fig. 1. There are various forms of tags and tag clouds. For instance, there are hashtags, data clouds, text clouds. A hashtag does not have to be a proper word or a phrase in some language. It can be any sequence of characters. Hashtags originated from tags and tagging popularized by Twitter. Hashtag was even chosen a "Word of the year 2012" by American Dialect Society (2013). Tag clouds can be built from hashtags as well. Data clouds or text clouds are specialized forms of tag clouds visualizing numerical data or word frequencies. For the rest of the paper we will use generic terms of a tag and tag clouds.

The first step in tag cloud creation is preparation of tags themselves: phrase selection, weighting, clustering (Fujimura et al., 2008; Lohmann et al., 2015; Nesi et al., 2016; Spyrou and Mylonas, 2016), etc. Methods of digesting the text and extracting the tags rest in text mining area and are beyond the scope of this paper. Here it is assumed that the set of tags is given and their rendering in two dimensions is studied. Tag clouds have been considered in the scientific circles for more than 10 years. In the early stage tag clouds could have been managed with direct researcher attention for better applicability and visual results. However, it is not possible for a mass application of web engineering and the user-dedicated content. Therefore, tag construction must be delegated to automatic tools tailored to the capability of browser clients. Tag clouds are used by web designers all over the world, often with poor results. The methods of tag cloud construction are often coined ad hoc, resulting in bad aesthetics or low usability. In this paper we analyze this problem and propose new methods to solve it for good aesthetics. A solution of the tag cloud construction problem must address at least three interdisciplinary challenges:

- Modeling and algorithmically controlling tag cloud aesthetics,

* Corresponding author.
E-mail addresses: jakub.marszalkowski@cs.put.poznan.pl (J. Marszałkowski), maciej.drozdowski@cs.put.poznan.pl (M. Drozdowski).

http://dx.doi.org/10.1016/j.engappai.2017.06.023
Received 16 January 2017; Received in revised form 28 May 2017; Accepted 26 June 2017
Available online 26 July 2017
0952-1976/

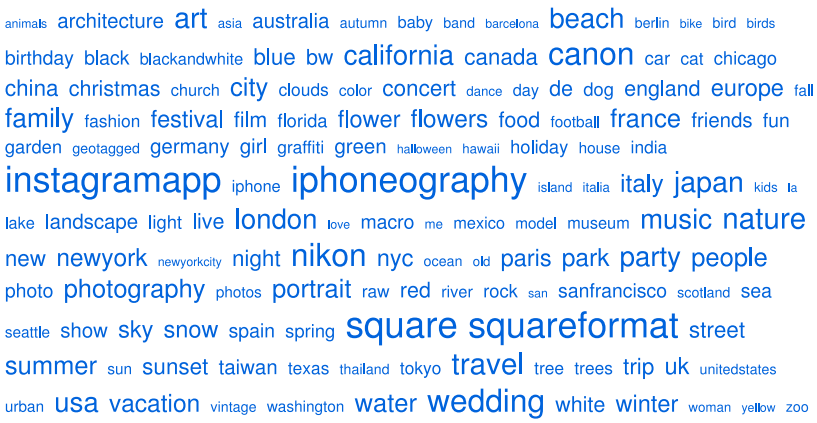

Fig. 1. Tag cloud from Flickr instance in Table 3.

- constructing the layout of tags by solving a 2D packing problem,
- advanced software engineering which meets soft real-time performance requirements and resource constraints of the browser platform.

The main contributions of the paper are as follows:

- algorithms for building tag clouds are surveyed and their taxonomy is proposed;
- usability studies are analyzed to identify areas of practical tag cloud application;
- the requirements and restrictions of tag clouds for browser exposition are analyzed resulting in formulation of the tag cloud construction problem as a combinatorial optimization problem with a dedicated irregular objective function;
- the invention of the objective function as a pivotal element linking the optimization process with tag cloud aesthetics;
- using rules of typography to control aesthetics;
- algorithms solving the problem are implemented, evaluated and selected for the target browser platform;
- proof of the concept that the role of a web designer may be automated by advanced optimization algorithms generating web pages online, on the client sides.

Overall, we solve the tag cloud construction for web exposition problem starting from the analysis of the context up to providing practical algorithms.

Further organization of this text is the following. In Section 2 taxonomy of tag clouds in general is proposed. Approaches, algorithms, design options and the choices taken in the past are surveyed. Requirements for tag clouds in web usage are determined in Section 3. Section 4 provides a formulation of the tag cloud construction problem as a combinatorial problem with constraints and irregular objective functions. Algorithms solving the problem are introduced in Section 5. These algorithms are tested for quality of solutions and conformance with the performance constraints of the browser platform. Results of the computational experiments are outlined in Section 6. The notations used in the paper are summarized in Table 2.

2. Related work survey

Although tag clouds seem to be a modern invention, their origins can be traced back at least to 1976 (Milgram and Jodelet, 1976). Early tag clouds history is outlined in Viégas and Wattenberg (2008). Around 2003 they gained a wide usage over the Internet. In 2006–2009 they became bloated, overused by many web-designers without considering whether they fit the purpose. Consequently, they were criticized and their application declined. Currently, a new generation of tag cloud approaches is proposed and tag clouds can be seen where they fit well. Thus, tag clouds seem to follow the hype cycle (Fenn and Raskino, 2008) and they slowly reach the productivity stage. Many approaches to the tag cloud construction have been used in the past. In this section we classify and analyze them, as well as, outline the results of the studies on tag cloud formation and usability.

2.1. Tag cloud taxonomy

There are design choices which determine appearance and usability of tag clouds. These design choices lead to tag cloud classification. In particular, the design decisions apply to:

1. Tag ordering rule. The options are: alphabetically, by importance, by context, randomly, packing-decided. The last means that tags may be reordered for better packing quality. Sorting by context means that tags are placed in groups connected, e.g., semantically, lexically, or in some other way.
2. Shape of the entire cloud. Possible options: rectangular, other regular (e.g. circular), irregular, given (e.g. given polygons, map borders used for visualization). Regular shapes may also have a ragged margin, which is often considered a typographical defect resulting from bad text justifying.
3. Shape of tag bounds. Options: rectangle, or character body. The former means that bounding boxes of the tags rendered in some given font are used. The latter means using the shapes of the characters in the given font. This allows for tighter tag alignment using free space around the letter bodies.
4. Tag rotation: none, free to rotate, allowed with limited degrees of freedom.
5. Vertical tag alignment. Options: sticking to the typographical baselines, limited by the algorithm properties (e.g. some tags are grouped), free—possibly leading to 2D packing, forced by the tag cloud background (e.g. a given heat map).

The consequences of the design decisions can be compared in Figs. 1 and 2 (cf. the design decisions outlined in Table 1). There are still further design-choices possible. For instance, it is possible to use differing colors or fonts (typefaces, sizes, weights and styles). In this paper we assume that fonts are determined in the tag preparation step (for example, chosen by the web designer), and hence given as input. Note that use of colors to distinguish tags may be a bad idea for users with color-impaired sight. Thus, we assume that tags are essentially monochromatic on a contrasting background (e.g. black on white).

2.2. State of the art

Since the very start tag clouds construction attracted interest of researchers. Kaser and Lemire (2007) experimented with two types of