The uncertain tag cloud

Manolis Wallace and Nikos Platis

Knowledge and Uncertainty Research Laboratory
Department of Informatics and Telecommunications
University of Peloponnese
Tripolis, Greece 22 100
Email: gav@uop.gr - Web: http://gav.uop.gr

***Abstract*—Tag clouds provide an excellent means of visualization of weighted semantic information. When, on the other hand, this information is not definitive but is rather accompanied by a measurable degree of uncertainty, conventional tag clouds are no longer suitable visualization tools. In this paper we extend the conventional approach to tag cloud generation and propose the utilization of the degree of opaqueness as a means to visualize the degree of certainty. In order to experimentally assess the efficacy of the proposed approach we have developed the corresponding software tools and have applied the conventional and proposed approached to tag visualization in a real life scenario of probabilistic data.**

Keywords—tag cloud; uncertainty; visualization

I. Introduction

The tag cloud originally came into prominence as a web navigation method, but more recently it has emerged as a very intuitive means of semantic information visualization.

A typical modern use of a tag cloud is the visualization of textual information, with frequency of term occurrence being used to determine the font size for each term. As a result, a user can acquire a rough understanding of the content of a large textual corpus without having to read it, simply by taking a quick look at an image. See for example Figure 1 where free text responses from 4000 individuals [5] are summarized in one brief picture. Using tag clouds to visualize semantic information is also quite common, with font size determined by some measure of importance.

For the tag cloud notion to be applicable, the depicted information needs to be complete and certain. Still, real life semantic information rarely is. Uncertainty is inherent in all aspects of human life and semantic information is certainly no exception. In this paper we propose the visualization of uncertain information using different levels of opaqueness. In order to run our experiments, but also in order to facilitate others in applying our approach, we have developed a suitable extension of a popular Java based tag cloud generator.

The rest of the paper is organized as follows: In section II we review related background such as information visualization, tag cloud generation, and types of uncertainty. Continuing, in section III we present our proposed approach using a simple example and in section IV we focus on the software tool that implements it. Finally, in section V we discuss further work needed in order to determine the optimal way to use the proposed approach and in section VI we list our concluding remarks.

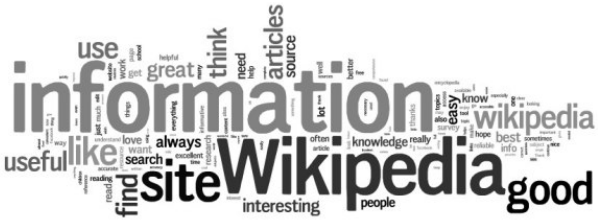

Fig. 1. Word cloud of open ended responses from the Wikipedia Readers Survey [6]

II. Related work

A. Information Visualization

Tag clouds are a characteristic tool of Information Visualization. Information Visualization uses graphical representations of data in order to enhance human cognition [11], to ease understanding of the data, to allow the viewer to form a mental model of it.

Very often raw data would be practically meaningless without a visual representation. Imagine a long time series of several stock values and try to reason about it just by looking at the numbers; it would be impossible. Now imagine a simple line graph of these stock values against time; within moments you would be able to discover the evolution of the price of each stock over time, the relative value of various stocks, etc.; you could even discern trends of the stock values which would otherwise require complex statistical analyses to reveal.

Information Visualization employs various graphical techniques in order to achieve its goals. One of its most important strengths is its ability to use different characteristics of the graphical representation in order to convey different properties (or dimensions) of the data; for example, the tags on a tag cloud could be colored differently if they belong to distinct categories; points on a diagram could depict many more properties apart from those assigned to their x and y coordinates by varying their size, color, and shape.

B. Tag clouds

Tag clouds are very common nowadays and can be found on numerous places, ranging from websites and blogs to scientific publications. Still, although most computer users know what to expect when they hear the term “tag cloud” no specific and universally acceptable definition exists for it. For example dictionary.com [9] defines a tag cloud as

> a visual representation of user-generated electronic tags or keywords that classify and describe online content, typically an alphabetical list or a grouping of words in different font sizes, as to show relative frequency or provide links to further information

which limits the term’s application to online content. Daniel Nations’ definition [10]

> A tag cloud is a box containing a list of tags with the most prominent or popular tags receiving a darker and bigger font than less popular tags.

is more inclusive, especially if one ignores the focus on popularity and interprets it more generically as an indicator of weight. But no constraints are put on which tags are to be included in a tag cloud or how their weight is to be mapped to font size and color. Also, this excludes cases in which color is not used jointly with font size to depict weight but independently to depict another information parameter, such as category, or set randomly.

Both of the above definitions exclude cases in which only the terms themselves are of interest and both size and color are set randomly to achieve a desired aesthetic effect as in Figure 2. And of course, as is to be expected in any situation for which a clear definition does not exist, there are those who consider Figure 2 to be a legitimate tag cloud and those who do not.

So, to avoid any ambiguity, we start by stating that in this work, we focus on tag clouds in the sense of pictures of tags in which font size is used to indicate a tag’s weight. The generation of such tag clouds is based on manually set maximum and minimum font sizes; the tag with the greatest weight is drawn using the maximum font size, the tag with the least weight is drawn using the minimum font size and tags with intermediate weights are drawn using intermediate font sizes, sometimes assigned linearly but most commonly assigned using a logarithmic scale which helps keep font sizes in reasonable ranges when the range of weights is large.

But even when considering only the above definition, there are numerous different approaches that can be followed in the development of a tag cloud. The font size to be used may be specified, but numerous other characteristics, such as font face and font style, tag orientation and tag placement, overall shape and collision mode, remain up to the designer/developer to determine. With respect to each one of these:

Font face and style. Although fonts are excellent candidates for the depiction of additional information in tag clouds, they have not received much attention in this way. A unique font is typically used throughout a tag cloud, with the selection of font face (eg. Arial, Tahoma, etc) and font style (eg. bold, italic, etc) determined solely based on aesthetics.

Tag orientation. In early tag clouds tag orientation was not considered at all, with all words printed horizontally. Lately a mix of horizontal and vertical tags is the most popular approach, with random angles also being used at occasions. The orientation of the tags does not carry any information and is selected either randomly or manually in order to achieve an aesthetically better result.

Fig. 2. “Thank you” in different languages [7]

Tag placement. Early tag clouds were simple lists of words printed in different sizes, starting on the top left corner of a box and continuing until all tags were listed. But a more recent, and much more intuitive, approach aims to place the tags of greater weight closest to the center of the tag cloud. Tags are first ordered in decreasing weight and then they are examined sequentially with each one placed as close to the centre as possible without overlapping with an existing word. The way to achieve this is by selecting for each tag a random angle and attempting to place it at that position with respect to the center of the cloud. If a collision with an existing tag is detected then the point of placement is moved, one step at a time, on an increasing spiral, until there is no collision. This is the approach that generates the elliptically shaped results found on many websites; figure [6] is such an example.

Overall shape. Early tag clouds were simply lists of printed words; later they became rectangle shaped, with words in each line distanced equally in order to achieve an alignment on both sides; now tag clouds are most commonly irregularly shaped, like clouds. In another recent approach that is gaining popularity, an image mask may be used to place the tags and create a visual impression; Figure 2 is such an example.

Collision mode. In all types of tag clouds there is one common characteristic: there are no overlaps between tags. Overlaps can be detected in two different ways: in the earlier and faster approach each tag is a rectangle and collisions are detected as intersections of rectangles, whilst in the most recent, more elegant but considerably slower approach each tag is turned into a bitmap mask and a pixel by pixel examination is performed. In the later approach smaller tags can be placed between the letters of larger words, leaving less white space and producing a more compact visual representation.

C. Uncertainty

Uncertainty is an inherent feature of human life and we are accustomed to dealing with it in almost any information that we are faced with. So much so, that we often do this intuitively

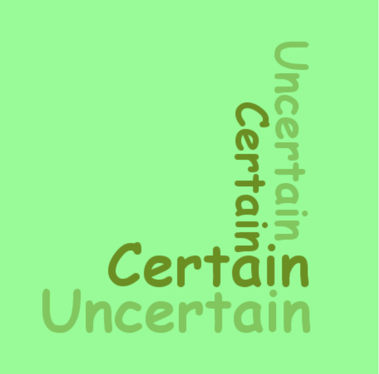

Fig. 3. The uncertain tag cloud concept

without even acknowledging its existence. More importantly, we typically consider the term “uncertainty” to refer to any situation in which something is not known certainly and accurately, not realizing that many heterogeneous types of information are included in that category.

We provide below a small and certainly not comprehensive list of types of uncertainty.

Probabilistic information. When it is not certain whether an event will occur, but the exact probability for that even is known. For example the chance that number 4 will be the result of a through of a fair dice is known to be exactly 1/6.

Possibilistic information. Similarly to the case of probabilistic information, it is not certain whether an event will occur; moreover, the exact probability is not known either, only some upper (plausibility) and lower (necessity) margins are known. For example the chance to win in a lottery, when the number of participants is not known.

Ambiguity. When the origin of the uncertainty does not lie within the data itself but rather within their interpretation. This is most common when using words to represent data.

Imprecision. Typically originating from the finite precision of a measuring tool. For example a person’s exact weight, when all we have to measure it with is a scale with an accuracy of complete kilos.

Vagueness. When the margins of a term’s meaning are not precisely and universally defined. Consider for example the meaning of the term “tall person” and the difficulty in categorizing some people as members or non-members of the set of tall people.

All of the above, and more, are very valid and very real cases of uncertainty. In this work, however, we only focus in cases that the uncertainty does not refer to the exact magnitude of things but rather to their very existence. Thus, probabilistic information is examined herein and addressed by our proposed approach, imprecision is not.

III. Visualization of uncertainty

As has been explain in the previous section, in this work we focus on types of uncertainty that are not correlated to the magnitude of things. Thus, these are cases in which the level of uncertainty is an independent variable of the data and should be visualized separately.

In the conventional tag cloud definition that we follow, magnitude is the weight of tags and is visualized by controlling the size of the font. Our proposal is to visualize the uncertainty of tags by controlling their opaqueness. This generates a very intuitive representation, with certain things being printed “normally”, absolutely improbable cases not depicted at all and intermediate cases drawn with varying levels of transparency. Figure 3 demonstrates the concept.

A. A real life setting

As an example, consider table I in which we summarize the upcoming year’s expected budget for a research group, together with each funding source’s probability. Projects A, B and C are running and next year’s budget is already secured. There is also a preliminary oral agreement with an industrial partner to implement project D during the next year, with the contract still pending but almost certain.

There are also proposals submitted to calls of different difficulties (projects E and F). And on top of that, there is the experience that some amount is typically secured during the course of every year, either by small ad-hoc projects or from departmental funds that remain unexploited at the end of the year and are distributed to the department members.

TABLE I. Expected budget for year 2016

Funding source	Amount in Keuros	Comments	P
ProjectA	50	Running project	1
ProjectB	100	Running project	1
ProjectC	150	Running project	1
ProjectD	30	Agreed project, to be contracted	0,9
ProjectE	100	Submitted proposal, easy call	0,5
ProjectF	200	Submitted proposal, very competitive call	0,1
Other	50	Additional funds typically attracted per year	0,8

Should we wish to depict this information in a tag cloud, we would be faced with the decision of how to visualize the different degrees of certainty related to each one of the table’s entries. One way would be to only depict most probable options, as shown in Figure 4. Alternatively, in Figure 5 we depict all options, choosing to hide the fact that we already know that some of these tags correspond to improbable situations.

In Figure 6 we incorporate uncertainty by weighing amounts proportionally to their probability. From an economic or risk analysis perspective this is the optimal approach, as what is depicted is the real economic value of each project at the present time. Still, from an information visualization point of view this is counter intuitive and misleading. For example consider project F which is depicted as small in scale. This is inaccurate in all cases as project F will either bring in a large amount or none at all.

The problem in Figure 6 stems from joining volume and uncertainty, which in our case are unrelated, in one visualization parameter. In order to overcome this, we propose presenting magnitude using font size and degree of uncertainty using transparency, as shown in Figure 7. Compared to the previous approaches, we observe that this visualization contains all of the information of the previous ones and communicates it in a straightforward manner.

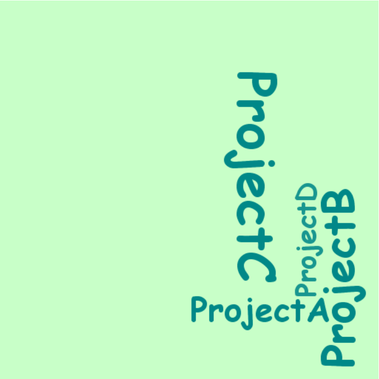

Fig. 4. Most probable funds for 2016

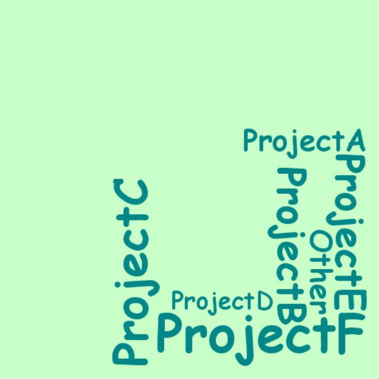

Fig. 5. All possible funds for 2016

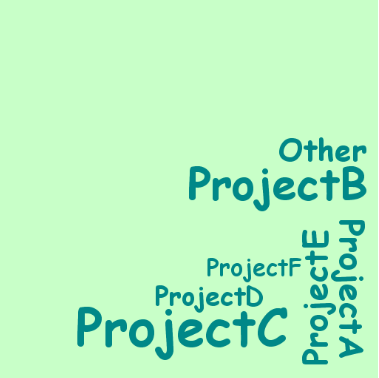

Fig. 6. Font sizes depict a combination of amount and probability

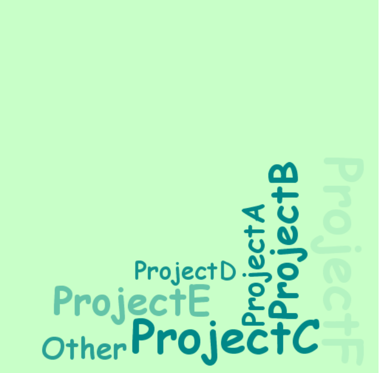

Fig. 7. The uncertain tag cloud of the 2016 budget

IV. The Tool in Use

Given the length of time that the notion of the tag cloud has been available, the wide range of its applications and, more importantly, the ambiguous way in which many of its parameters are determined, it is only natural that there are various software applications [12][13][14][15][16] and libraries [8][17][18][19] available for the creation of tag clouds. And whilst they have in common the use of word frequencies or some other weight to determine font size, other choices such as exact placement of words, text colors, overall size and shape, etc., differ considerably among them.

Our proposed approach to the representation of uncertainty is not related to these choices, and therefore it may be combined with any of the abovementioned software. In order to experimentally demonstrate the effectiveness of the proposed approach we have chosen to extend the Kumo - Java Word Cloud library [8] allowing for the utilization of different degrees of opaqueness for each word, as determined by the user's input.

Kumo is a popular starting point for those who wish to develop their own tag cloud generator, due to the fact that it is open source and to the flexibility it provides:

- It supports word overlap checks at either word level (with each word corresponding to one rectangle), or pixel level (with the shape, size and position of each letter examined).
- It is customizable in image size.
- It is customizable in maximun and minimum font size.
- It supports predetermined, custom or random color palettes.
- It supports the generation of tag clouds of different directions.
- It supports the generation of tag clouds that resemble any user determined shape.
- It is available as an open source Java tool and can therefore easily be integrated with other Java applications

These features and characteristics are all preserved in the new version we have developed.

We should note that Kumo also has some notable weaknesses, the most important one being that it is focused more on the appearance of the tag cloud than the validity of its content; when the inclusion of a word will prevent the tag cloud from assuming the intended shape, Kumo opts to omit the word, thus generating a beautiful but semantically invalid tag cloud.

This shortcoming does not prohibit us from using it to develop our proof of concept. Thus, all of the figures presented in section II-C have been generated using this software.

Similarly to the Kumo library, our finalized tool and accompanying libraries will be made freely available under a GPLv3 licence in our GitHub account.

V. What does it mean?

A powerful characteristic of the conventional tag cloud is that everyone comprehends its meaning with a first glance, mainly because the meaning of the font size is always the same. Similarly, we need to make sure that a single meaning for the degree of opaqueness is used in all cases, or that the meaning of the opaqueness is always the most intuitive one.

In the example of section III we used the degree of opaqueness to describe probabilistic information. But we saw in section II-C that there are many more and fundamentally different types of information that are grouped under the general umbrella of uncertainty, some of which may be visualized using the uncertain tag cloud approach.

More importantly, it is worth noting that the applicability of the proposed approach is not necessarily limited to the visualization of uncertainty. Seen from a more generalistic point of view, the opaqueness can be used as an indicator of the degree to which a specific tag is in context; in this paper's example, tags are in context when they correspond to actual funding. Thus, we can envision the uncertain tag cloud approach and software used to indicate semantic relevance to a given concept, geographic distance from a point of interest, time distance from a moment in time etc.

Consider for example a tag cloud on a tourist information site depicting the most visited cultural sites in or near Athens, with font size corresponding to the annual number of visitors. The Acropolis will certainly be in it, Sounio at a 1h 15" drive will probably be in it, but what about Olympia at a 3 hour drive? Different levels of opaqueness could certainly facilitate the accurate and intuitive visualization of this information.

As part of our future work we aim to compile a comprehensive list of types of information, situations and scenarios when the approach proposed herein may be applied.

VI. Conclusions

Tag clouds are a very common form of visualization for lexical information. In this paper we focused on uncertain information and proposed the utilization of opaqueness to indicate the degree of certainty.

We developed the corresponding software and successfully applied our approach on probabilistic data. Yet, we feel there is a wider range of situations in which it will be useful; we plan to investigate these situations as part of our future work.

References

[1] O. Kaser and D. Lemire, *Tag-cloud drawing: Algorithms for cloud visualization.* arXiv preprint cs/0703109, May 2007.

[2] M. Halvey and M.T. Keane, *An Assessment of Tag Presentation Techniques*, poster presentation at WWW 2007, 2007

[3] A.A. Alola, M. Tunay and V. Alola, *Analysis of Possibility Theory for Reasoning under Uncertainty* International Journal of Statistics and Probability; Vol. 2, No. 2; 2013

[4] J. Sinclair and P. Cardew-Hall, *The folksonomy tag cloud: when is it useful?* Journal of Information Science 34.1, 2008, pp. 15–29.

[5] Research:Wikipedia Readership Survey 2011/Results https://meta.wikimedia.org/wiki/Research:Wikipedia_Readership_Survey_2011/Results

[6] M. Pande, *Word Cloud of open ended responses from Wikipedia's readers survey*, via Wikimedia Commons

[7] K. Latimer, *Saying Thank YouGlobally*, November 20, 2012 http://newhydepark.tipsfromtown.com/2012/11/20/saying-thank-you-globally/

[8] Kenny Cason, Kumo - Java Word Cloud, http://kennycason.com/posts/2014-07-03-kumo-wordcloud.html

[9] Dictionary.com, "tag cloud," in *Dictionary.com Unabridged.* http://dictionary.reference.com/browse/tag%20cloud

[10] D. Nations, *What is a "tag"? – What is a "tag cloud"?*, http://webtrends.about.com/od/glossary/g/tag_def.htm

[11] C. Ware, *Information Visualization, Perception for Design*, 3rd edition, Morgan Kaufmann, 2012.

[12] Wordle http://www.wordle.net/

[13] Tagcloud Generator http://www.tagcloud-generator.com/

[14] Word Cloud Generator http://www.jasondavies.com/wordcloud/

[15] Tagxedo http://www.tagxedo.com/

[16] TagCroud http://tagcrowd.com/

[17] OpenCloud http://opencloud.mcavallo.org/

[18] WordCram http://github.com/danbernier/WordCram/

[19] d3-cloud http://github.com/jasondavies/d3-cloud/