Tag clouds and retrieved results: The CloudCredo mobile clustering engine and its evaluation

Stefano Mizzaro, Luca Sartori, and Giacomo Strangolino

Department of Mathematics and Computer Science
University of Udine
Udine, Italy
mizzaro@uniud.it, sartori.uni@gmail.com, delleceste@gmail.com

Abstract. We discuss the use of tag clouds as a way of visualizing the results of a clustering search engine. We briefly present a specific tag cloud approach and its implementation in the CloudCredo prototype. Then we describe an experimental user study aimed at demonstrating that tag cloud visualization is: (i) as effective as classical tree like visualization; and (ii) particularly effective on small screen devices. Towards the aim (i), we compare CloudCredo with a similar system, Credino; towards (ii), in the experiment the two systems are compared on iPhone and iPad, two similar devices differing mainly in their size. Results, although preliminar, support the hypotheses.

Keywords: Clustering, Mobile devices, Tagcloud, Evaluation

1 Introduction

On the Web, there is a growing number of *clustering search engines*, namely search (or, more often, meta-search) engines that present the retrieved documents organized in clusters: similar documents are grouped together under a meaningful label; clusters are organized hierarchically (i.e., clusters are divided into sub-clusters, and so on) and usually shown in a tree-like manner; and the end user can browse the retrieved results by focusing on specific clusters. Some examples of these systems are: Yippy (formerly known as Clusty and Vivísimo) www.yippy.com or CREDO credo.fub.it.[1] Even classical search engines like Google show some signal of a clustering approach, although they are still much more oriented towards the classical ranked list.

The cluster approach seems particularly adequate and effective for mobile devices, since it allows to use the limited screen space in a more effective way. This approach has been proposed and evaluated for the CREDO system, and its mobile versions Credino and SmartCREDO [2,3]. Indeed, mobile search engines are an important and hot research topic: as it is well kwon, several statistics

[1] At the time of writing CREDO is not available.

show that Internet traffic in general, and queries to search engines in particular, generated by means of mobile devices are quickly increasing. It is foreseen that by 2015 there will be more mobile users than desktop Internet users.

However, the classical tree-based visualization of document clusters is not the only possibility. In this paper we propose a *tag cloud* based visualization that, in our opinion, has the potential to be particularly effective on small-screen mobile devices. Our aim is twofold:

- to understand if the tag cloud visualization is effective;
- to understand if it is particularly effective on mobile device small screens.

The paper is organized as follows: Sect. 2 defines tag clouds and motivates our approach; Sect. 3 presents CloudCredo, a mobile clustering engine implementing the tag cloud approach, and recalls Credino, a companion system used in the evaluation; Sect. 4 describes the user study that we performed to experimentally evaluate the tag cloud effectiveness.

2 Tag clouds

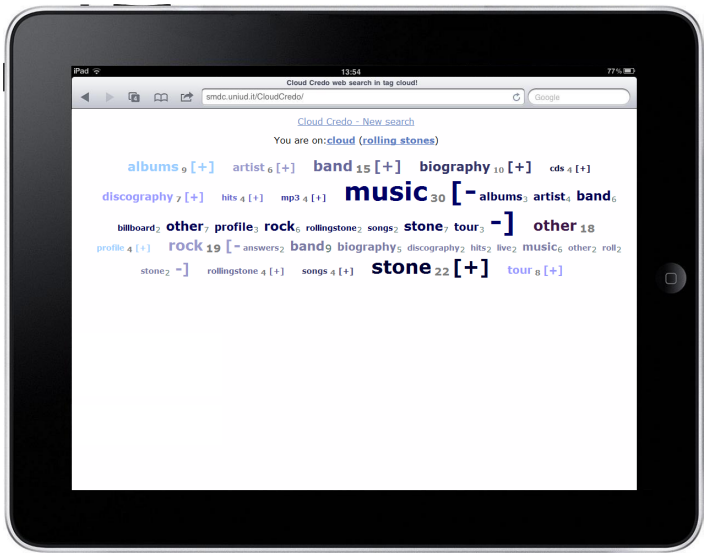
A tag cloud (or word cloud) is a set of terms organized spatially and graphically (in terms of fonts and colors) to visually highlight the most important terms. Tag clouds are very common: they are being used quite often on the Web, to show the tags used to annotate web resources, to summarize the main topics of a Web site, and so on. There are several kinds of tagclouds, that can differ for the selection of terms, the graphical aspect, and the auxiliary information shown (like a count of each term frequency in the original text); a description can be found at `en.wikipedia.org/wiki/Tag_cloud`.

As mentioned above, we propose to use a tag cloud to show the label of the clusters. The rationale for this approach is that a tag cloud can show the same labels and use less space than the classical tree-like visualization, although admittedly in a less organized way. Moreover, not only the cluster labels are shown as a tag cloud, but the labels are clickable, and can be expanded into sub-clusters (as in the tree like visualization). Also, we specifically tailor mobile devices, and we are interested in studying the effectiveness of the tag cloud clustering approach when screen space is limited.

3 Credino and CloudCredo

We build on Credino system [2], implemented with the aim of porting the CREDO clustering engine on a mobile device (a PDA was used in the original paper [2], but we slightly adapted it to more recent devices like the iPhone). Figure 1(a) shows a screenshot of Credino on an iPhone. On the basis of Credino, we implemented CloudCredo, that visualizes the same clusters as Credino by means of a tag cloud. Figures 1(b) and (c) show the screenshot of CloudCredo on an iPhone and an iPad. Credino and CloudCredo are both meta-search engines on CREDO, therefore they both show exactly the same cluster hierarchy,

(a) Credino (b) CloudCredo on iPhone



(c) CloudCredo on iPad

Fig. 1. Credino (a) and CloudCredo (b) on an iPhone with the query [rome]: the cluster "city" has been expanded and the subclusters are shown. CloudCredo (c) on an iPad, with the query [rolling stones], and the clusters "music" and "rock" expanded. Scale factor are different: with the same scale, the iPad would be almost twice bigger.

just visually different. As can be seen in Figures 1(b) and (c), our tag cloud implementation exploits both colors and size, and each cluster also shows the number of documents in it. The tag cloud implementation, similarly to the classical hierarchical tree one, allows to expand a category into subcategories, by clicking on the "`[+]`" sign close to the tag (and to compact it by clicking on "`[-]`"). Both Credino and CloudCredo are Web applications that can be used by any standard Web browser; on iPhone and iPad they adapt smoothly to the portrait/landscape orientation of the device.

We are not alone in proposing to use tag clouds to show the retrieved results; the Quintura search engine `www.quintura.com/` does exactly that. Our approach is slightly different, though, since: (i) our tags/clusters can be expanded into subtags/sub-clusters; and (ii) we specifically target mobile devices in this work.

CloudCredo is available at `smdc.uniud.it/CloudCredo`; the version of Credino used in the experimental evaluation described below is at `credino.dimi.uniud.it/`. The two systems, being based on CREDO (see Footnote 1), are not available at the time of writing.

4 Experimental evaluation

We performed a user study towards the two aims stated at the end of Section 1. These can be translated into the following experimental hypotheses: (i) CloudCredo is as effective as Credino; and (ii) CloudCredo effectiveness turns out to be high in particular on small screens.

4.1 Experimental design

We used the two systems Credino and CloudCredo in our evaluation. We also used an iPhone and an iPad: since the two devices are very similar, the main (if not only) difference being their size, we try in this way to single out the effect of size. Thus, our experiment has two independent variables:

- device, or size (iPhone and iPad);
- system (Credino and CloudCredo).

48 participants, recruited in our university, were involved in our study. Each participant was asked to perform 4 tasks. The tasks were built by starting from the most frequent queries on Google Mobile `www.google.com/intl/en/press/zeitgeist2008/`: we selected 4 of them and built 4 simulated work task situations [1] around them. Figure 2 shows task 1, translated from Italian to English, as given to the user. To have a more controlled environment, we specified the initial query. To limit learning effects, we relied on a Graeco-Latin square design: each subject performed her 4 tasks on the four system/device combinations, in a different order.

As dependent variables, we measured both objective user effectiveness and subjective user satisfaction. User effectiveness was measured as a linear combination of: the success in finding the appropriate page (a binary value in $\{0, 1\}$),

Task 1

- **Description:** Imagine that you are going to visit a friend in Rome and therefore you want to find some information about cultural events (e.g., exhibitions and concerts) that will take place during your stay in town.
- **Task:** Retrieve two different pages.
  A page is relevant if it provides the date, time, and location of an event taking place in Rome during the next 30 days. Pages discussing an event in a general way, without specifying the above data, will be not relevant.
- **Other instructions:** Start with the query [rome events]

Fig. 2. The first task used in the experiment.

the speed (computed on the basis of time needed and normalized into the $[0, 1]$ range), and the confidence the user had to have performed her task correctly (again normalized into $[0, 1]$). We defined three different combinations of these three factors, with different weights; however, there was no difference among the three combinations. In the following we measure effectiveness E as

$$E = \text{success} * (2/3 * \text{speed} + 1/3 * \text{confidence})$$

(if success is 0, then E is 0 as well; speed is more important than confidence).

User satisfaction was measured by means of questionnaires: participants filled in a questionnaire after each task completion, and one final questionnaire as well. Questionnaires collected, by means of Likert scales, data about:

- difficulty of the task;
- difficulty of using the system;
- adequacy of the system to the device.

We combine these three values into a single satisfaction one S' by taking their average, normalized onto $[0, 1]$:

$$S' = 1/3 * \text{task\_d} + 1/3 * \text{system\_d} + 1/3 * \text{adequacy}.$$

We also take into account other two questionnaire items that, as a control, asked whether the participant preferred the other system or device. The final satisfaction S was computed by slightly changing S' to take these into account.

We adopted the usual procedures of a laboratory testing: each subject was briefed and trained, she filled in a first questionnaire with some demographics data, then she started the four task-questionnaire iteration, and finally filled in the last questionnaire. We also ran a pilot test, that confirmed the choice of the four tasks and allowed to estimate the maximum time allowed for each task.

4.2 Results

The collected demographics show that participants were either university students (45 out of 48) or just graduated searching for a job. They had good — and

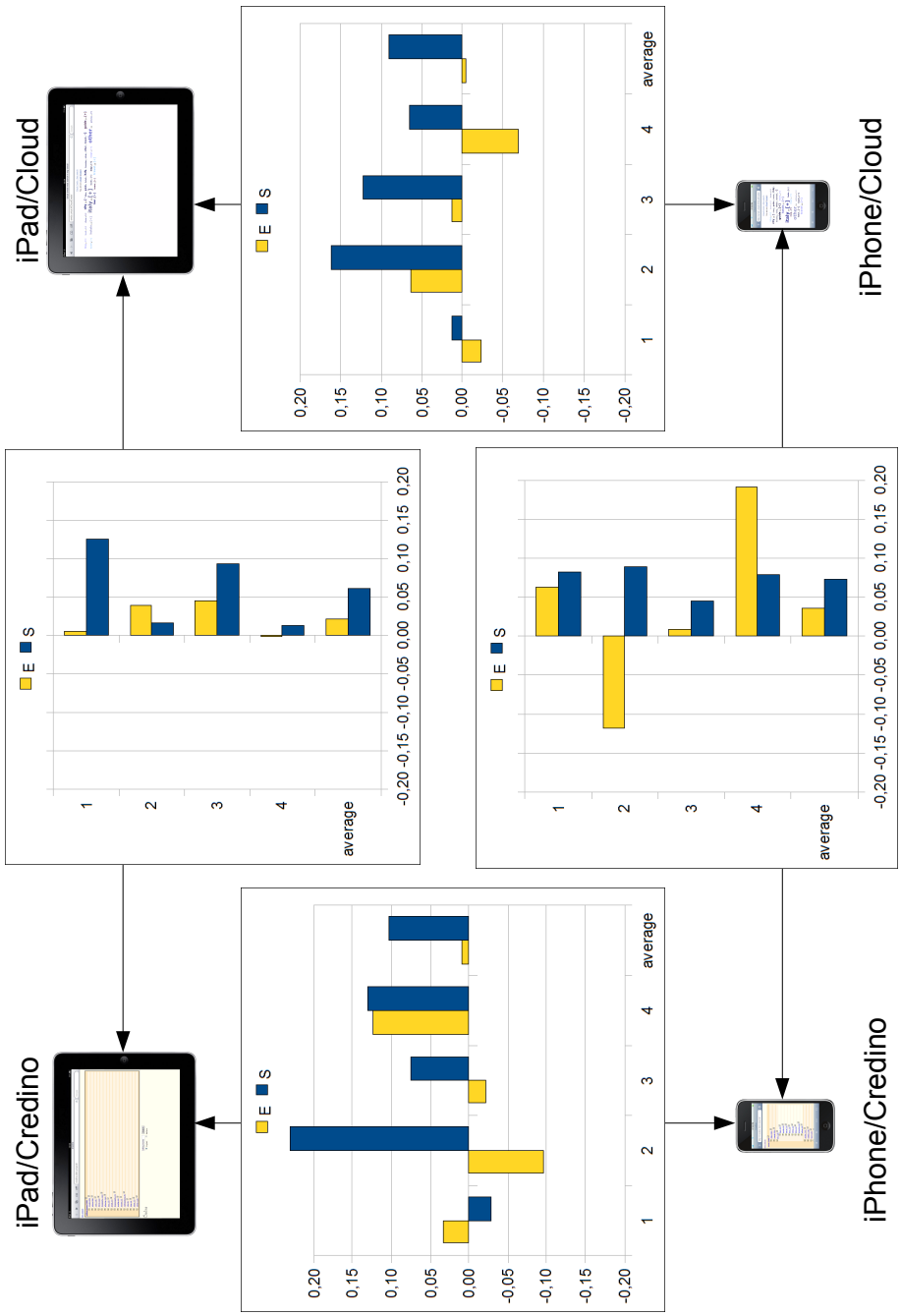


Fig. 3. Overall results.

homogeneous — knowledge of computers, Web search, and mobile devices. All of them were aware of iPhone and iPad devices. Nobody had used a clustering engine before.

Figure 3 shows the overall results. The four device/system combinations are shown in the corners of the figure; the four charts show the differences — in both S and E — for the single tasks and averaged on the four tasks. The bars are oriented towards the best device/system combination, e.g., the leftmost vertical bar shows that iPad/Credino had a higher effectiveness E than iPhone/Credino on task 1.

By analyzing the figure we can understand that:

- Since all the "average" bars point towards right, on average, CloudCredo had both a higher E and a higher S than Credino.
- Since all the average bars point towards up (with a single exception, the rightmost E bar, which is anyway very small in absolute value), on average, the iPad device had both a higher E and a higher S than iPhone.
- Combining the previous two points, iPad/Cloud was the most effective and most preferred combination.
- The above considerations seem stronger for S, which has longer bars.
- We can see that the above results hold for most of the single tasks as well: there are only 6 bars on specific tasks that disagree with the average bar (out of 32 possibilities).
- Also, on the single tasks, E and S are often in agreement, although in 6 out of 16 cases they are not.
- Although we were interested in showing that the tag cloud visualization was as effective as the tree-like one, these results are a first cue that it is even more effective and preferred. However, there is almost no statistical significance on the differences. On E, according to the Mann-Whitney-Wilcoxon test, the only statistically significant difference (at the 0.05 level), is on task 4 between iPhone/Credino and iPhone/Cloud (the longest horizontal E bar in figure). Statistical significance is slightly higher on S: although most of the differences are not significant, the preference of iPad/CloudCredo to iPad/Credino is significant at the 0.05 level, according to the Mann-Whitney-Wilcoxon test.
- Therefore, the two visualization approaches can be considered equivalent, with a slight preference for the tag cloud one. This confirms the first hypothesis.
- Turning to the second hypothesis, the figure shows that the average difference is slightly higher at the iPhone level than at the iPad one. There is no statistical significance for this result, however, also because there are quite high variations over the single topics (i.e., bars on the top chart are often very different from the corresponding bars on the bottom chart — see, for example, the striking difference on the E value on task 2). Thus we can only say that there is a slight indication of the particular effectiveness of the tag cloud approach on small screens, also on the basis of the results in [2,3] that showed how the clustering approach of Credino is more effective on small screens than on large ones.

5 Conclusions

We have proposed a tag cloud based approach to the visualization of the retrieved results by a clustering search engine. Our experimental study on two prototypes supports the hypotheses that tag clouds are an effective visualization alternative, especially on small screen mobile devices. The second point is particularly critical, since we do not have a statistically significant proof of it. We do not have any contrary evidence, though; this, combined with the results of previous studies [2, 3] makes indeed interesting the option of using a tag cloud based approach on mobile devices, although further evidence should be found.

The experimental design needs some further remarks. The usual user study performed in information retrieval aims at demonstrating that a new version of some system reaches higher effectiveness and/or user satisfaction than some baseline. Our experimental study was somehow different from this classical setting, since we were interested in showing that an alternative system (actually, visualization approach) is as effective as a classical one.

Although the results of our user study are positive, they are preliminary: we used four tasks only, and the user population is quite homogeneous. Therefore, a first and obvious future work direction is to repeat the experiments with a higher number of tasks and with a different, and perhaps more heterogeneous, user population. Also, a more sophisticate experimental design can help to prove the second hypothesis. A last direction is to implement native applications for iPhone/iPad (and Android as well) of CloudCredo and Credino: this would allow a more effective interaction and a better user experience.

References

1. P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3):paper no. 152+, 2003.
2. C. Carpineto, A. Della Pietra, S. Mizzaro, and G. Romano. Mobile Clustering Engine. In *Proceedings of the 28th European Conference on Information Retrieval, London, UK*, volume 3936 of *Lecture Notes in Computer Science*, pages 155–166. Springer, 2006.
3. C. Carpineto, S. Mizzaro, G. Romano, and M. Snidero. Mobile information retrieval with search results clustering: Prototypes and evaluations. *Journal of the American Society for Information Science and Technology*, 60(5):877–895, 2009.