Improving Tag Clouds with Ontologies and Semantics

Antonio M. Rinaldi
DIS - Dipartimento di Informatica e Sistemistica - Universitá di Napoli Federico II
Via Claudio, 21 - Napoli, Italy 80125
CeRICT scrl - Centro Regionale Information Communication Technology
Via Traiano, 1 - Benevento, Italy 82100
Email: antoniomaria.rinaldi@unina.it

***Abstract*—Tag clouds are visual representations of a set of terms which represent several document dimensions. Social and collaborative systems have greatly increased the popularity of this type of visualization but several problems arise from their knowledge base structures. In this paper we propose a novel strategy to improve tag clouds with ontological and semantic information. Our methodology is based on a general knowledge base to extract additional terms and relations. These information are combined with statistics to enhance tag clouds visualization and improve their possible application in user-based systems.**

***Keywords*-Tag clouds; ontologies; semantic networks; WordNet**

I. Introduction

The extremely rapid growth of user centered information on the internet due to the development of “Social Web” applications requires of novel approaches to help users during their information searches and browsing. More and more people use tagging services and enjoy them as discovery tools. Indeed, tagging is simple, it does not require a lot of thinking and it is very useful to find relevant objects. People tag pictures, videos, and other resources with a couple of keywords to easily retrieve and share them in a later stage.

There are several ways to aid users in these tasks and in the last years new techniques have been proposed. One of these approaches is based on the creation of tag clouds. They show a set of terms in which text features (e.g. size, color, weight) are used to represent relevant properties among words and collected documents. They can be arranged along different visual features aggregation as (i) a tag for the frequency of each item; (ii) a global tag cloud where the frequencies are aggregated over all items and users; (iii) a cloud contains categories, with size indicating number of subcategories.

Tag clouds can be used for basic user-centered tasks [1] as: *Search* - Locating (or determining the absence of) a specific target or alternative target; *Browsing* - Casually exploring the cloud without a specific target or purpose; *Impression Formation and Impression Presentation* - The cloud can be scanned to get a general idea about a subject; *Recognition or Matching* - Recognizing the entire cloud as data which describes a subject.

Tag clouds arise from collaborative tagging paradigm [2], [3] used in social software website as flickr (http://www.flickr.com), delicious (http://www.delicious.com), Technorati (www.http://technorati.com) and Bibsonomy (http://www.bibsonomy.org). In these systems, users annotate contents with free keywords (tags) defining associated metadata without any need to use existing, pre-defined and authoritative indexing structures; this classification system is called folksonomy [4]. Folksonomies have a high impact on user tasks and are in strong contrast with other forms of terms classifications (e.g. thesauri and ontologies).

This visualization tool implies less cognitive and physical workload than thinking of a search tag that defines the thematic field one likes to explore and entering it into the search field [5]; for example, after having found an initial tag and associated resources users can start browsing using tags or make use of related tag lists. Even if tag clouds have been shown to help users get a high-level understanding of the data and to support people in casual exploration [1], the completely free choice of tags entails several problems for users. For example, it is hard to have a full impression of tags used in the whole system, users are often dealt with general linguistic problems related to folksonomies [6], [7], structured ways of exploration are hardly provided and user interfaces of folksonomy systems often fail to support users in finding appropriate search tags and creating efficient queries for discovering interesting contents. Moreover, as discussed in [8], [9], if visible tags are selected only by their usage frequency, there might be a problem of high semantic density, which means that very few topics and related prominent tags tend to dominate the whole visualization and less important items fade out [10].

In this paper we propose novel algorithms, techniques and metrics to improve tag clouds with ontological and semantic information. Our approach is based on ontologies automatically extracted from a general knowledge base and a metric to measure tag features by means of semantic and statistical properties.

The paper is organized as follows: in section II some related works are presented together with the difference with our approach; the proposed strategy, algorithms and metrics are discussed in section III; in section IV evaluation methodology and several results are presented; eventually, discussions on our approach and conclusions are in section V.

II. Related Works

Several studies have been presented in literature to add more information to folksonomies and enhance tag visualization in order to improve the use of tag clouds. A model to merge ontologies and social networks using tagging mechanism is presented in [11]. The relations among objects arise from graph transformations of annotation structure in order to obtain a tag co-occurrence graph including the co-occurrence counts for each pair of tags. A user interface approach called Semantic Cloud is described in [12]. The system allows users to explore the tag space of a folksonomy system within a hierarchical structure of semantically arranged tag clouds representing different topics and their subtopics. Grahl et al. [7] and Gemmell et al. [13] present algorithms to build hierarchical structures from folksonomies to provide a more effective browsing or personalized navigation, respectively. Several approaches [9], [14] have been proposed to measure tag similarity using statistics. An interesting approach to construct semantic networks on the basis of tag cooccurrences with the goal of comparing the network structures of folksonomies is in [15]. The same authors analyzed similarities between tags and documents in order to enrich semantic aspects of social tagging. An interface for information searching task using tag clouds has been presented in [5]. The authors point out that tag clouds, as visual summaries of content, satisfy all the roles mentioned in [1], and they observed that the process of scanning the cloud and clicking on tags is easier than the formulation of a search query. Kaser and Lemire [16] optimize the usability of tag clouds trying to establish a relation between similar tags. From their point of view, similarity does not mean that the tags represent the same semantic concept, but rather that they were used to describe the same document. Schrammel et al. [17] evaluated the effects of semantic arrangement versus alphabetical and random arrangement of tags in tag clouds. They observed that a semantically clustered tag cloud with randomly arranged tags yields an improvement for specific queries and aids in directing the users attention towards tags with a smaller size. Clustering algorithms were applied to gather semantically similar tags. In [8] the k-means algorithm was applied to group semantically similar tags. Li et al. [18] supported a large scale social annotations browsing based on an analysis of semantic and hierarchical relations. In [19] the authors investigate ways to support semantic understanding of collaboratively generated tags. They conducted a survey on practical tag usage in Last.fm, an on-line music community. Based on the results, they propose a visualization named TagClusters, in which tags are clustered into different semantic groups and the visual distance represents the semantic similarity between tags. Zubiaga et al. [20] presented a methodology to obtain and visualize a cloud of grouped tags based on the use of SOMs, and language models. Semantic representation should ideally involve associating user interests with appropriate URIs, thus moving folksonomy user profiles closer to the Semantic Web and moving the agenda of using Semantic Web technology to organize collectively assembled information characteristics of Web 2.0 [21]. Semantically-Interlinked Online Communities (SIOC) is an ontology that provides a foundation for semantically representing user activities in blogs and forums [22]. To facilitate representing tags with URIs, Meaning Of A Tag (MOAT) was developed as a framework to help users manually select appropriate URIs for their tags from existing ontologies [23]. Specia and Motta [24] investigated on reusing of existing ontologies to link tags automatically with pre-crafted concepts and relations.

In our paper we propose a different strategy based on ontologies dynamically extracted from a general knowledge base used to smooth problems related to floksonomies and we use an overall metric to combine statistical and semantic information for tag clouds visualization.

III. The Proposed Strategy

In this section we describe our strategy putting in evidence the novelty of our approach and the used techniques. We argue that several problems highlighted in the previous sections and related to folksonomies and tag clouds structure can be relaxed using ontologies [25] and metrics to compute tag clouds elements.

We use WordNet [26] as general knowledge base and, because some tasks are accomplished using WordNet properties, before reporting and detailing each phase, it's useful to introduce some considerations about the WordNet structure, so we can better understand our algorithm and techniques.

All information in WordNet is arranged using linguistic properties. The basic unit is the synset, a logic set of words related through the synonymy property. Each synset is a concept in WordNet. All the synsets are related to the others by pointers that represent linguistic properties. Two kinds of relations are represented: *lexical* and *semantic*; lexical relations hold between word forms while semantic relations hold between word meanings.

In our approach the terms in the analyzed document are used as input to build domain ontologies (i.e. semantic networks) extracted from WordNet. Then, these ontologies are intersected in order to have new terms related to the document context; the new recognized terms are added to the tag cloud and the visual features of the cloud are computed using an ad hoc metric later described.

A. Tag cloud enhancement task

In this step we analyze the terms in the considered document to add new ones. This task is accomplished by an innovative algorithm to build dynamically domain ontologies, represented as semantic networks (SN), using WordNet.

The semantic network is built starting from the synset that represents a concept identified by a term in the document. We then consider all the component synsets and construct a hierarchy, only based on the hyponymy property; the last level of our hierarchy corresponds to the last level of WordNet one. After this step we enrich our hierarchy considering all the other kinds of relationships in WordNet. Based on these relations we can add other terms in the hierarchy obtaining an highly connected semantic network.

The algorithm to extract the semantic network is described in pseudo-code in Table I.

TABLE I
SN GENERATION ALGORITHM

```
//--------------------------------------------------------
// SN creation algorithm
//
// INPUT: Main_Synset: represents the considered synset
//
// OUTPUT: Synset_List: the list returned from the function.
//         It contains all SN synsets
//--------------------------------------------------------
Synset_List   CreateSN (Main_Synset)
{
  Add Main_Synset to a Synset_List
  Load from Wordnet the Category_terms of Main_Synset
  Add founded synsets to Synset_List
  While (Synset_List<>EOF)
  Do {
     Load from Wordnet all hyponyms of all synsets
     in Synset_List
     Add founded synsets to Synset_List
  }
  While(Synset_List<>EOF)
  Do {
     Load from Wordnet all synsets linked to all synsets
     in Synset_List using all linguistic properties
     (counting hyponymy and hypernymy out)
  }
  return Synset_List
}
```

Due to the polysemy property (i.e. the capacity for a sign or signs (e.g., a word, phrase, etc.) to have multiple meanings) we built several SNs considering the same representative but the number of common synsets between a semantic network which represents a concept out of our context of interest with other semantic networks is very low as shown in the system log files. Anyway, terms wrongly added will have a poor visibility in the tag cloud due to the metric used to compute visual features described in the next section.

An example of SN is shown in Figure 1.

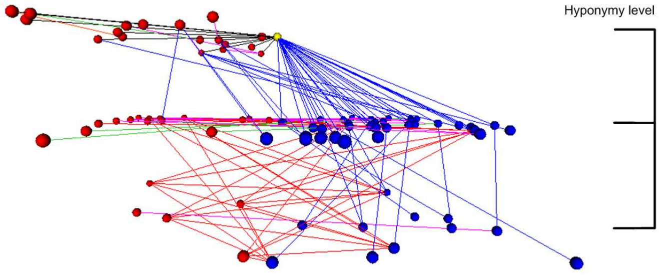

Fig. 1. A SN example - Car (Sense1)

At the end of this task, we have a list of terms which represents the analyzed document. This list is composed by the words in the analyzed document together with new terms from the semantic networks intersection.

B. *The used metric*

In our approach we propose a novel technique to combine statistical information and semantic properties of terms in a document introducing a measure to take into account the weight of a single term in the document itself.

To calculate the relevance of a term we assign a weight to each one in the tag cloud considering the polysemy property, that can be considered as a measure of the ambiguity in the use of a word, if it can assume several senses.

Thus we define as *centrality* of the term i as:

$$\varpi(i) = \frac{1}{poly(i)} \tag{1}$$

poly(i) being the polysemy (number of senses) of i.

As an example, the word *car* has five senses in WordNet, so the probability that it is used to express a specific meaning is equal to $1/5$. We argue that those words have only one meaning strongly characterize the expressed concept.

We define our metric also considering statistical information by means of term-weight functions because they should favor terms that are representative of the document, but should also discriminate between the documents in a collection. Based on these considerations, we propose a term weighting approach based on compound normalized weights with three factors [27]:

- **Term frequency**: is the number of occurrences of a term in a document;
- **Document frequency**: is the number of documents within the global information space in which the term appears;
- **Document size factor**: compensates for high term frequencies of terms in large documents.

We are now in a position to introduce our metric:

$$M_{i,k} = \frac{(a+(1-a)(TF_{i,k}/TF_{max,k}))(\log N/n_i)\varpi_i}{\sqrt{\sum_{i\in k}(a+(1-a)(TF_{i,k}/TF_{max,k})(\log N/n_i)(\varpi_i))^2}} \tag{2}$$

k being the list of terms related to the k-th document, *i* being the i-th term, $TF_{i,k}$ being the term frequency of *i* in *k*, $TF_{max,k}$ being the maximum term frequency in *k*, *N* being the total number of documents in collection, n_i being the number of documents to which the term *i* is assigned, ϖ_i being the centrality of *i*, *a* being a smoothing term whose role is to damp the contribution of the second term which may be viewed as a scaling down of TF by the largest TF value in k. The basic idea is to avoid a large swing in the normalized $TF_{i,k}$ from modest changes in $TF_{i,k}$. The value of *a* is set to 0.5 [28].

This formula give us statistical information about the analyzed document and the whole collection but, using the term centrality, we have a more accurate definition of the role of the considered term in the document. These semantic and statistical information are shown in tag cloud visualization by tag size. We explicitly point out that we are considering a collection of documents to perform our metric. This is a real scenario because many web sites which use tag clouds work with their communities and own documents.

Subject	Domain	Yahoo Category	Doc
Mars	Astronomy	Directory>Science>Astronomy>Solar System >Planets >Mars	22
	Mythology	Directory>Society and Culture>Mythology and Folklore >Mythology>Greek>Gods and Goddesses>Ares(Mars)	5
Davis	Music	Directory>Entertainment>Music>Artists>By Genre>Jazz >By Instrument>Trumpet>Davis, Miles (1926-1991)	12
	Sport	Directory>Recreation>Sports>Tennis>Tournaments>Davis Cup	11
Jaguar	Animal	Directory>Science>Biology>Zoology>Animals, Insects, and Pets >Mammals>Cats>Wild Cats>Jaguars	7
	Car	Directory>Recreation>Automotive>Makes and Models >Jaguar	19
	Sport	Directory>Recreation>Sports >Football(American) >Leagues >National Football League(NFL) >Teams >Jacksonville Jaguars	13
Apache	Computer	Directory>Computers and Internet>Software >Internet>World Wide Web>Servers>Unix>Apache	18
	Helicopter	Directory>Government>Military>Aviation>Helicopters>AH-64 Apache	9
Lincoln	History	Directory>Arts>Humanities>History>U.S. History>By Subject>Presidency>Presidents>Lincoln, Abraham (1809-1865)	15
	Car	Directory>Recreation>Automotive>Makes and Models >Ford	21

TABLE II
TEST SET EXAMPLE

IV. EVALUATION

We use a general *document collection* to evaluate our approach. We built the document collection by means of interaction with the directory service of the search engine Yahoo. The directory service provides the category referred to each Web page. The tag clouds generated from the document collection are analyzed by a questionnaire asked to a group of 50 users (MSc students and Ph.D. students of information science).

We compare common document term frequency counting (**STC**) and our new techniques (**ITC**) for tag cloud generation.

We use this methodology to build our document collection to have a complete evaluation of the different components of our metric and give us a support for user opinions understanding. The whole collection has been used during the evaluation task and in Table II a portion of it is shown together with some examples of its organization.

An example of generated tag cloud is in Figure 2

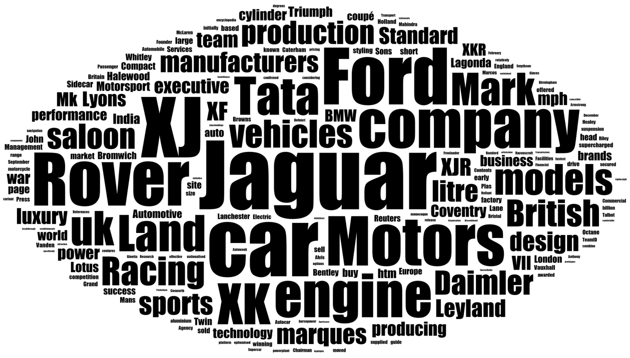

Fig. 2. Jaguar (car) tag cloud

We follow a methodology presented in [29] to evaluate our techniques. This method works with two sets of statements: measure expectations about a service category in general (E) and statements to measure perceptions (P) about the category of a particular service. Each statement is accompanied by a 7-point scale ranging from *strongly disagree* (1) to *strongly agree* (7).

Specific studies show that it is possible to adopt this methodology for measuring effectiveness of information systems [30]. Moreover, we consider several indicators [31], [32], [33], [34] to measure different dimensions of our approach, namely: (i) perceived usefulness (**PU**), (ii) perceived ease of use (**PEU**), (iii) trust in the information system (**TIS**), and (iv)perceived enjoyment (**PE**).

The documents and the related tag clouds have been randomly assigned to users and evaluated using the indicators.

Figure 3 show the evaluation results in terms of mean (**M**) and standard deviation (**SD**).

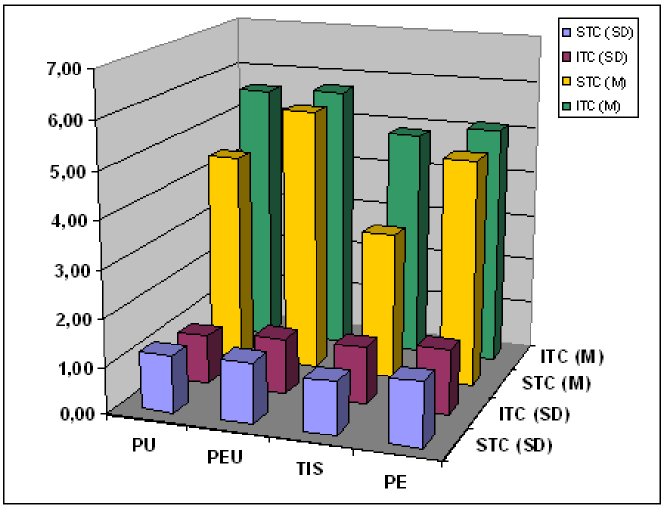

Fig. 3. Evaluation results

The generated tag clouds can be used for several purpose in concert with the issues highlighted in Section I.

V. Conclusions

The social experiences on the web involve more accurate tool for representing and sharing information. In this context tag clouds are a powerful and representative implementation. On the other hand the user centered approach in the new vision of internet and the needs of effective methodologies for data and application cooperation and understanding encourage the use of formal knowledge representations as ontologies and semantics.

In this paper we propose a novel strategy to combine such kind of information to improve tag cloud visualization. Our strategy is general and the generated tag clouds can be used for searching, browsing or representing documents. Evaluations results are promising and interesting issues seem to be increased using our approach as "serendipity", a term often used [2] referring to possible unexpected findings during browsing tags.

Actually we are investigating on the use of other semantic properties and more efficient metrics to measure the relatedness among document terms, tag clouds and folksonomies; moreover, other visual features can be used to combine these information and improve the quality of data visualization.

References

[1] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen, "Getting our head in the clouds: toward evaluation studies of tagclouds," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ser. CHI '07. New York, NY, USA: ACM, 2007, pp. 995–998.

[2] A. Mathes, "Folksonomies - Cooperative Classification and Communication Through Shared Metadata." Dec. 2004. [Online]. Available: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

[3] T. Hammond, T. Hannay, B. Lund, and J. Scott, "Social bookmarking tools (i): A general review," *D-Lib Magazine*, vol. 11, no. 4, April 2005.

[4] T. V. Wal, "Folksonomy coinage and definition," Website, March 2007. [Online]. Available: http://vanderwal.net/folksonomy.html

[5] J. Sinclair and M. Cardew-Hall, "The folksonomy tag cloud: when is it useful?" *Journal of Information Science*, vol. 34, no. 1, pp. 15–21, 2008.

[6] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *J. Inf. Sci.*, vol. 32, no. 2, pp. 198–208, Apr. 2006.

[7] M. Grahl, A. Hotho, and G. Stumme, "Conceptual clustering of social bookmarking sites," in *7th International Conference on Knowledge Management (I-KNOW '07)*. Know-Center, Sep. 2007, pp. 356–364.

[8] Y. Hassan-Montero and V. Herrero-Solana, "Improving tag-clouds as visual information retrieval interfaces," in *InScit2006: International Conference on Multidisciplinary Information Sciences and Technologies*, 2006.

[9] G. Begelman, P. Keller, and F. Smadja, "Automated Tag Clustering: Improving search and exploration in the tag space," in *Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006*, Edinburgh, Scotland, 2006.

[10] M. A. Hearst and D. Rosner, "Tag clouds: Data analysis tool or social signaller?" in *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, ser. HICSS '08. Washington, DC, USA: IEEE Computer Society, 2008.

[11] P. Mika, "Ontologies are us: A unified model of social networks and semantics," *Web Semant.*, vol. 5, no. 1, pp. 5–15, Mar. 2007.

[12] H. Aras, S. Siegel, and R. Malaka, "Semantic cloud: An enhanced browsing interface for exploring resources in folksonomy systems," in *Workshop on Visual Interfaces to the Social and Semantic Web (VISSW2010), IUI2010*, 2010.

[13] J. Gemmell, A. Shepitsen, B. Mobasher, and R. Burke, "Personalizing navigation in folksonomies using hierarchical tag clustering," in *Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery*, ser. DaWaK '08. Springer-Verlag, 2008, pp. 196–205.

[14] K. Fujimura, S. Fujimura, T. Matsubayashi, T. Yamada, and H. Okuda, "Topigraphy: visualization for large-scale tag clouds," in *Proceedings of the 17th international conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 1087–1088.

[15] C. Cattuto, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme, "Network properties of folksonomies," *AI Commun.*, vol. 20, no. 4, pp. 245–262, 2007.

[16] O. Kaser and D. Lemire, "Tag-cloud drawing: Algorithms for cloud visualization," in *Proc. WWW 2007 Workshop on Tagging and Metadata for Social Information Organization*, Banff, Canada, 2007.

[17] J. Schrammel, M. Leitner, and M. Tscheligi, "Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches," in *Proceedings of the 27th international conference on Human factors in computing systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 2037–2040.

[18] R. Li, S. Bao, Y. Yu, B. Fei, and Z. Su, "Towards effective browsing of large scale social annotations," in *Proceedings of the 16th international conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 943–952.

[19] Y.-X. Chen, R. Santamaría, A. Butz, and R. Therón, "Tagclusters: Semantic aggregation of collaborative tags beyond tagclouds," in *Proceedings of the 10th International Symposium on Smart Graphics*, ser. SG '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 56–67.

[20] A. Zubiaga, A. P. García-Plaza, V. Fresno, and R. Martínez, "Content-based clustering for tag cloud visualization," in *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, ser. ASONAM '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 316–319.

[21] T. Gruber, "Collective knowledge systems: Where the social web meets the semantic web," *Web Semant.*, vol. 6, no. 1, pp. 4–13, 2008.

[22] U. Bojrs, J. G. Breslin, A. Finn, and S. Decker, "Using the semantic web for linking and reusing data across web 2.0 communities," *Web Semant.*, vol. 6, no. 1, pp. 21–28, 2008.

[23] A. Passant, "Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data," in *Proceedings of the Linked Data on the Web (LDOW2008) workshop at WWW2008*, 2008.

[24] L. Specia and E. Motta, "Integrating folksonomies with the semantic web," in *Proceedings of the 4th European conference on The Semantic Web: Research and Applications*, ser. ESWC '07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 624–639.

[25] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquis.*, vol. 5, no. 2, pp. 199–220, 1993.

[26] G. A. Miller, "Wordnet: a lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[27] G. Salton and C. Buckley, "Readings in information retrieval," K. Sparck Jones and P. Willett, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, ch. Term-weighting approaches in automatic text retrieval, pp. 323–328.

[28] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[29] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "Servqual: A multiple-item scale for measuring consumer perceptions of service quality," *Journal of Retailing*, vol. 64, no. 1, pp. 12–40, 1988.

[30] L. F. Pitt, R. T. Watson, and C. B. Kavan, "Service quality: a measure of information systems effectiveness," *MIS Q.*, vol. 19, no. 2, pp. 173–187, 1995.

[31] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.

[32] B. Kim and I. Han, "The role of trust belief and its antecedents in a community-driven knowledge environment," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 5, pp. 1012–1026, 2009.

[33] C. S. Lin, S. Wu, and R. J. Tsai, "Integrating perceived playfulness into expectation-confirmation model for web portal context," *Inf. Manage.*, vol. 42, no. 5, pp. 683–693, 2005.

[34] H. Heijden, "User acceptance of hedonic information systems," *MIS Q.*, vol. 28, no. 4, pp. 695–704, 2004.