Mining Social Media for Enhancing Personalized Document Clustering

Chin-Sheng Yang(✉) and Pei-Chun Chang

Department of Information Management, and Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Chung-Li, Taiwan, ROC
csyang@saturn.yzu.edu.tw, s996202@mail.yzu.edu.tw

Abstract. Social media is nowadays an excellent platform for gathering user intelligence for supporting business intelligence applications. Social tagging system (aka. folksonomy) is a critical mechanism for collaboratively creating, organizing and managing the wisdom of crowds. The knowledge gained from social tagging system should be tremendous assets for conducting and improving various business intelligent applications. Consequently, the purpose of this study is to examine the values of folksonomy on an important business intelligent task, namely personalized document management. Specifically, we employ Delicious, a pioneered social bookmarking service, to construct a statistical-based thesaurus which is then applied to support personalized document clustering. According to our empirical evaluation results, social tagging system indeed improve the quality of the statistical-based thesaurus in comparison with that constructed on the basis of a general-purpose search engine in generating personalized document clusters.

Keywords: Social media · Business intelligence · Social tagging · Social bookmarking · Personalized document clustering

1 Introduction

Social media is nowadays the most popular platform that allows the creation and exchange of user generated content [15]. According to the research results by the Pew Research Center [24], over 70 % of internet users use social media sites as of January 2014. Another report by eMarketer [12] reveals that, by the end of 2013, 163.5 million people in U.S.-more than two-thirds of internet users-will be social media users. Moreover, Facebook, the global leading social networking service provider, has 1.35 billion monthly active users as of the third quarter 2014 [13]. 4.5 billion "Likes" were generated and 4.75 billion pieces of content was shared daily as of May 2013. These statistics indicate the social appeal associated with social media and user-generated content and the value of acquiring information from social media to facilitate the development of novel and the improvement of existing products and services.

Various social media websites, such as wikis (e.g., Wikipedia), blogs and microblogs (e.g., Twitter), media sharing (e.g., YouTube, Flickr), social news (e.g., Digg, Reddit), social bookmarking (e.g., Delicious, CiteULike), and social networking (e.g., Facebook, Google+), have been established. The knowledge (aka "wisdom of

F.F.-H. Nah and C.-H. Tan (Eds.): HCIB 2015, LNCS 9191, pp. 185–196, 2015.
DOI: 10.1007/978-3-319-20895-4_18

crowds") gained from social media sites can not only meet the objectives of businesses offering them but also help the development of novel and effective services that are better tailored to users' needs. In this study, we focus on analyzing a specific mechanism, i.e., social tagging system (aka folksonomy), commonly supported by numerous social media sites, e.g., YouTube, Flickr, Delicious, etc., for enhancing the effective of personalized information management. A folksonomy is a system of classification [29] which allows users to attach self-defined keywords (or tags) to describe resources [21], [27]. Folksonomy generally consists of a set of users, a set of self-defined tags, a set of resources, and a set of tag assignments (i.e., a set of user-tag-resource triple relationships) [8]. Semantically, tags in a folksonomy reflect users' collaborative cognition on information. They can reveal both the users' behavior and resources' properties [34].

The knowledge gained from folksonomy is valuable for supporting various applications, such as Web page classification [1], recommendation [22, 37], and information retrieval [3, 6]. In this study, we attempt to apply the wisdom of crowds of folksonomy to a novel document management task, namely personalized document clustering. Specifically, we adopt the CAC technique proposed by Yang and Wei [36] as our underlying personalized document clustering algorithm. The CAC technique takes into consideration a user's categorization preference (expressed as a list of anchoring terms) and subsequently generates a set of document clusters from this specific preferential perspective. Furthermore, the CAC technique exploits the world wide web as an information source to construct a statistical-based thesaurus, which then serves to expand the set of anchoring terms which is then applied to represent the source documents and then performs clustering to generate document clusters in accordance with the input preferential context (i.e., initial set of anchoring terms provided by the target user). Alternatively, we want to understand the effectiveness of folksonomy, in comparison with a general-purpose search engine (i.e., Google in Yang and Wei's study), on constructing a statistical-based thesaurus for supporting personalized document clustering. We select delicious (https://delicious.com/), a leading social bookmarking site, as the folksonomy for our social-tagging-based CAC technique (ST-CAC). We also conduct some experiments to evaluate the effectiveness of the ST-CAC technique and its benchmark approaches.

The remainder of this paper is organized as follows. Section 2 reviews existing document clustering techniques relevant to this study. In Sect. 3, we describe the detailed design of the proposed ST-CAC technique. Subsequently, we depict our experimental design and discuss important evaluation results in Sect. 4. Finally, we conclude with a summary and some future research directions in Sect. 5.

2 Literature Review

Document clustering entails the automatic organization of a large document collection into distinct groups of similar documents that reflect general themes hidden within the corpus [23, 32]. The documents in the resultant clusters exhibit maximal similarity to those in the same cluster and, at the same time, share minimal similarity with documents in other clusters. However, according to the context theory of classification, document clustering behaviors of individuals not only involve the attributes (including

contents) of documents but also depend on who is performing the task and in what context [2, 7, 17]. As a result, document clustering is an intentional act that should reflect individuals' preferences with regard to the semantic coherency or relevant categorization of documents [26] and should conform to the context of a target task under investigation.

Most of existing document clustering techniques are anchored in document content analysis. The overall process of a content-based document clustering technique generally comprises three main phases: feature extraction and selection, document representation, and clustering [14, 32, 33]. The purpose of feature extraction and selection is to extract and select from the target document corpus a set of representative features to represent the documents in the document representation phase. Subsequently, the clustering phase applies a clustering technique to group the target documents into distinct clusters.

Feature extraction begins with the parsing of each source document to produce a set of nouns and noun phrases and exclude a list of prespecified "stop words" that are non-semantic-bearing words. Subsequently, representative features are selected from the set of extracted features. Feature selection is important for clustering efficiency and effectiveness, because it not only condenses the size of the extracted feature set, but also reduces the potential biases embedded in the original (i.e., nontrimmed) feature set [25, 35]. Commonly used feature selection metrics include: TF, TF × IDF, and their hybrids [4, 19].

On the basis of a particular feature selection metric, the k features with the highest selection metric scores then are selected to represent each source document in the document representation phase. Based on the chosen representation scheme, each document is described in the k-dimensional space and represented as a feature vector. Commonly employed document representation schemes include binary (presence or absence of a feature in a document), within-document TF, and TF × IDF [4, 19, 23, 25, 32].

In the final phase of document clustering, source documents are grouped into distinct clusters on the basis of the selected features and their respective values in each document. Common clustering approaches include partitioning-based [4, 9, 19], hierarchical [11, 25, 30, 32], and Kohonen neural network [18, 20, 25].

As mentioned, content-based document clustering techniques rely on an objective feature-selection metric (e.g., TF or TF × IDF) that merely considers document content. As a result, existing content-based techniques generate for all users an identical set of document clusters from a given document collection and, thus, is unable to support personalized document-clustering. In response to the limitation of existing content-based document clustering techniques, prior research has proposed several extended approaches that might support personalized document clustering. For example, Deogun and Raghavan [10] propose a user-oriented document clustering technique that considers only document relevance to user queries. Kim and Lee [16] propose a semi-supervised document clustering technique to improve clustering effectiveness. Their approach essentially is a hybrid one that considers not only content similarity but also a user's perception of the document similarity using a relevance-feedback mechanism. Wei et al. [32] instead propose a personalized document clustering (PEC) approach to support personalization in document categorization.

In addition to the contents of the documents to be clustered, the PEC approach includes a target user's partial clustering as input, because it reflects his or her categorization preference. Last, Yang and Wei [36] propose a context-aware document-clustering (CAC) technique that takes into consideration a user's categorization preference (expressed as a list of anchoring terms) and subsequently generates a set of document clusters from this specific preferential perspective.

The abovementioned extended document clustering techniques in some degree can support the desired personalized document clustering task. Accordingly to Yang and Wei's study [36], the CAC technique outperforms other extended approaches in terms of supporting personalized document clustering. Thus, we adopted the CAC technique as the underlying algorithm for personalized document clustering. The CAC technique adopt a general-purpose search engine (i.e., Google) to construct a statistical-based thesaurus which serves as the basis for generating a set of document clusters which fits the categorization preference of a specific user. In this study, we adopt social media (more specifically, social tagging system) as an alternative information source for statistical-based thesaurus construction. The rational is that the information in folksonomy has been processed by crowds and reflects users' collaborative cognition. Such collaborative wisdoms should be better in supporting personalized document clustering.

3 Proposed Method

The context-aware document-clustering (CAC) technique, proposed by Yang and Wei [36], takes into consideration a user's categorization preference (expressed as a list of anchoring terms) and then generates a set of document clusters from this specific preferential perspective. For example, given a set of research articles related to "data mining," a person interested in developing new data mining techniques may prefer document categories anchored on the techniques under discussion and thus provides some anchoring terms as classification analysis, clustering analysis, association rules, sequential patterns, and so on. On the other hand, another person, who is working on data mining techniques to real world business applications, may prefer a different set of categories based on the application domains involved (e.g., banking, retailing, health care, telecommunications, etc.). Given the set of user-provided anchoring terms which represent the specific user's categorization preference, the CAC technique first constructs a statistical-based thesaurus and subsequently expands the given set of anchoring terms by adding their relevant terms. The expanded set of anchoring terms is adopted as the representative features for performing personalized document clustering.

The major difference between the CAC technique and our extended social-tagging-based CAC technique (ST-CAC) is the way of constructing statistical-based thesaurus. As shown in Fig. 1, the ST-CAC technique consists of five main phases: (1) feature extraction and selection; (2) statistical-based thesaurus construction; (3) anchoring term expansion; (4) document representation; and (5) document clustering. The detailed design of each phase is described in this section.

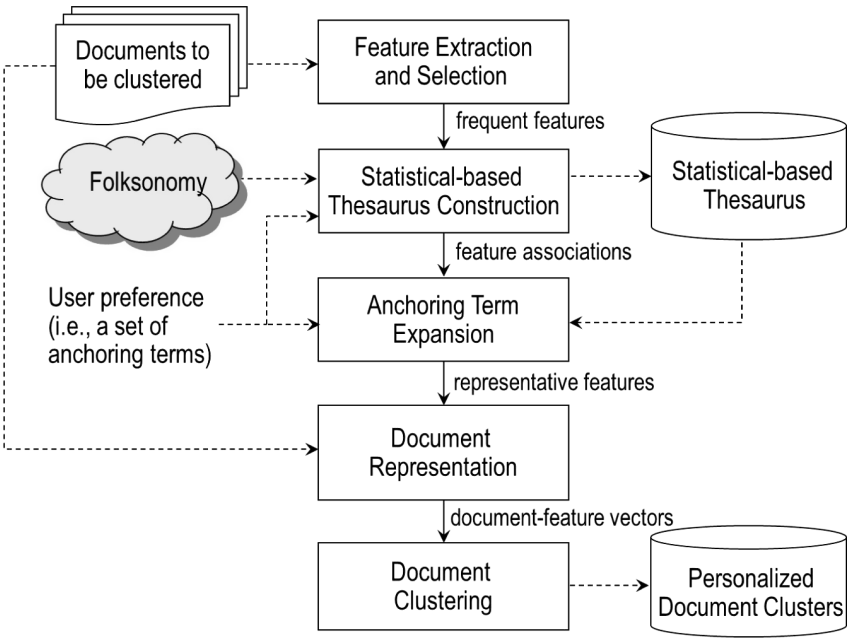

Fig. 1. Overall process of the ST-CAC technique

3.1 Feature Extraction and Selection

This phase aims at extracting and selecting a set of meaningful features (specifically, nouns and noun phrases) from the target document corpus. We adopt the part-of-speech (POS) tagger developed by Brill [5] to syntactically tag each word in the target documents and then employ Voutilainen's approach [31] to implement a noun-phrase parser for extracting noun phrases from each tagged document. Furthermore, we remove features that infrequently appear in the target document corpus. Particularly, we only retain those features whose document frequency (*df*) is no less than a prespecified threshold δ_{DF}.

3.2 Statistical-Based Thesaurus Construction

The purpose of this phase is to automatically construct a statistical-based thesaurus that will be used for expanding the user-provided anchoring terms. We adopt the folksonomy of Delicious website as the corpus for constructing a statistical-based thesaurus. Folksonomy generally is consisted of a set of users (U), a set of self-defined tags (T), a set of resources (R), and a set of tag assignments $A \subseteq U \times T \times R$ (i.e., a set of user-tag-resource triple relationships). A bookmark in Delicious website is a triple (u, T_{ur}, r) with $u \in U$, $r \in R$, and a set of tags $T_{ur} = \{t \in T \mid (u, t, r) \in A\}$.

For each anchoring term q_i pertaining to the categorization preference of a target user and every feature f_j representative to the target document corpus, the proposed

ST-CAC technique calculates the relevance weight between q_i and f_j by the pointwise mutual information (PMI) measure [28] as follows:

$$rw_{q_i,f_j} = \log_2\left(\frac{p(q_i \wedge f_j)}{p(q_i)p(f_j)}\right) = \log_2\left(\frac{N \times \text{hits}(q_i \wedge f_j)}{\text{hits}(qi)\text{hits}(f_j)}\right), \quad (1)$$

where $rw_{qi,fj}$ denotes the relevance weight between q_i to f_j, $p(query)$ is the probability that *query* (i.e., q_i or f_j) been used as a tag to annotated some resources (i.e., $p(query) = |R_{query}|/|\mathrm{R}|$, where R_{query} is the set of resources which are annotated with tag *query*), N is total number of resources in the folksonomy (i.e., $|\mathrm{R}|$), and hits(*query*) is the number of resources which are annotated with the tag *query* (i.e., $|R_{query}|$).

We extend the standard PMI measure by incorporating the number of users $U_t = \{u \in U \mid (u, t, R) \in A\}$ who use the tag t to annotated at least one resource. A tag commonly used to annotate same resources should have higher weight than those infrequently adopted. Accordingly, the weight PMI is defined as:

$$weighted\_rw_{q_i,f_j} = log_2\left(\frac{W\_N \times sum\_u(q_i \wedge f_j)}{sum\_u(q_i)sum\_u(f_j)}\right), \quad (2)$$

where $sum\_u(query)$ is the summation of number of users who use tag *query* to annotate some resources (i.e., $sum\_u(query) = \sum_{query \in Rquery} \left|U_{query}\right|$) and W_N is the total number of tag assignments (i.e., $|A|$). We employ the weighted PMI measure for our proposed ST-CAC technique.

3.3 Anchoring Term Expansion

On the basis of the constructed statistical-based thesaurus, this phase expands a given set of anchoring terms AT by including additional relevant terms. An anchoring term q_i in AT is expanded with a set of terms E_{qi} whose relevance weights, measure by weighted PMI values, to q_i need to be greater than a prespecified threshold α. The expanded set of anchoring terms $RF = \left(\bigcup_{q_i \in AT} E_{q_i}\right) \cup AT$ is formed for the succeeding document clustering task.

Because RF consists of the anchoring terms originally provided by the target user and relevant terms expanded from the anchoring terms, the importance of the terms in RF should not be identical when they are used to represent each document to be clustered. Accordingly, a TF × IDF-like scheme is adopted to estimate the weight of each expanded term f_j (i.e., in RF but not in AT) as:

$$w_j = \sum_{q_i \in ET_j} rw_{q_i,f_j} \times \log\left(\frac{|AT|}{|ET_j|} + \varepsilon\right), \quad (3)$$

where $ET_j \subseteq AT$ is the set of anchoring terms that expand f_j and ε is a small positive value to avoid the log component being 0. On the other hand, if $f_i \in AT$, w_j is the largest weight across all expanded terms derived previously.

3.4 Document Representation

Subsequently, each document to be clustered is represented using the expanded set of anchoring terms *RF*. ST-CAC employs the weighted TF × IDF scheme for document representation. Specifically, each document d_l is described by a feature vector $\overrightarrow{d_l}$ as:

$$\overrightarrow{d_l} = \langle v_{l1} \times w_1, v_{l2} \times w_2, \ldots, v_{lm} \times w_m \rangle, \tag{4}$$

where m is the total number of terms in *RF*, v_{lj} is the standard TF × IDF value of f_j in d_l, and w_j is the weight of the term f_j in *RF*.

3.5 Document Clustering

Finally, the target documents are grouped into distinct clusters on the basis of the expanded set of anchoring terms (i.e., *RF*) and their respective representation values in each document. ST-CAC adopts the hierarchical clustering approach (specifically, the HAC algorithm with the cosine measure for the similarity estimation between two documents and the group-average link method for similarity measurement between two clusters) as the underlying clustering algorithm.

4 Empirical Evaluation

This section reports our empirical evaluation of the proposed ST-CAC technique using a traditional content-based document clustering technique and the CAC technique as performance benchmarks. In the following, we discuss the evaluation design (including data collection and evaluation criteria), parameter tuning experiments, and important evaluation results.

4.1 Data Collection

The collection of document corpus for our evaluation purpose consists of 434 research articles related to information systems and technologies that were collected through keyword searches (e.g., XML, data mining, robotics) from a scientific literature digital library website (i.e., CiteSeer, http://citeseerx.ist.psu.edu/). For each article in our literature corpus, only the abstract and keywords were used in this evaluation study.

To evaluate the effectiveness of a personalized document clustering technique, we need to categorize our literature corpus from different users' preferential perspectives. We developed a system to collect individuals' preferred clustering for the literature corpus. Each experimental subject was asked to subjectively categorize the entire

Table 1. Summary of subjects' categories for the literature corpus

	Number of clusters (i.e., anchoring terms)	Number of documents in a cluster
Maximum	67	125
Minimum	10	1
Average	26.12	16.64

literature corpus manually on the basis of his/her own preference. After clustering, the subject was asked to assign a label for each category. These category labels are then considered as the set of anchoring terms of the subject which will be used as the input to the ST-CAC technique. A total of 33 subjects accomplished the manual clustering of the literature corpus. According to the self-reported estimates of the subjects, each subject spent a minimum of eight hours performing manual document clustering. A summary of the document categories generated by the subjects is provided in Table 1.

4.2 Evaluation Criteria

We employ cluster recall and cluster precision [25], defined according to the concept of associations, to measure the effectiveness of the ST-CAC technique and its benchmark techniques. An association refers to a pair of documents that belong to the same cluster. Accordingly, the cluster recall (*CR*) and cluster precision (*CP*) from the viewpoint of a subject u_a is defined as:

$$CR = \frac{|CA_a|}{|T_a|} \quad \text{and} \quad CP = \frac{|CA_a|}{|G_a|}, \tag{5}$$

where T_a is the set of associations in the categories manually produced by the subject u_a, CA_a is the set of correct associations that exists in both the clusters generated by a document-clustering technique and the categories produced by u_a, and G_a is the set of associations in the clusters generated by the document-clustering technique.

To address the inevitable trade-offs between cluster recall and cluster precision, precision/recall trade-off (PRT) curves are employed. A PRT curve represents the effectiveness of a document clustering technique with different intercluster similarity thresholds.

4.3 Parameter Tuning

We randomly select 10 users from the 33 subjects to determine the appropriate value of each parameter involved in the three document clustering techniques (i.e., a traditional content-based approach, the CAC approach, and our proposed ST-CAC approach) examined. The overall clustering effectiveness of each technique in the tuning experiments is calculated by averaging the cluster recall and cluster precision obtained from the ten subjects.

The traditional content-based document clustering (TCC) approach involves the parameter of number of features (k) for document representation. We range k form 200 to 2000 in increments of 200 and obtain the best performance when k is equal to 2,000. On the other hand, both CAC and ST-CAC techniques include the parameters δ_{DF} (the threshold to remove infrequent features in the feature extraction and selection phase) and α (the threshold to determine whether a term should be expanded in the anchoring term expansion phase). We first investigate α from 1 to 10 in increments of 0.5. The best values of α for CAC and ST-CAC are 2.5 and 2 respectively. Subsequently, we examine δ_{DF} from 3–10 in increments of 1 and get the best δ_{DF} values of 10 and 9 for CAC and ST-CAC respectively.

4.4 Comparative Evaluation Results

Using the parameter values determined previously, we evaluate the effectiveness of the ST-CAC technique and its benchmark techniques. In this experiment, all of the 33 subjects are used for evaluation purpose. The comparative evaluation result is shown in Fig. 2. The proposed ST-CAC technique achieves better clustering effectiveness than do the TCC and CAC techniques. Moreover, the CAC technique also outperforms the TCC technique. These results suggest that both ST-CAC and CAC techniques indeed have the ability to generate personalized document clusters according to the target user's personalized preference expressed as a set of anchoring terms. Furthermore, using social media for statistical-based thesaurus construction has better performance than that constructed from a general-purpose search engine.

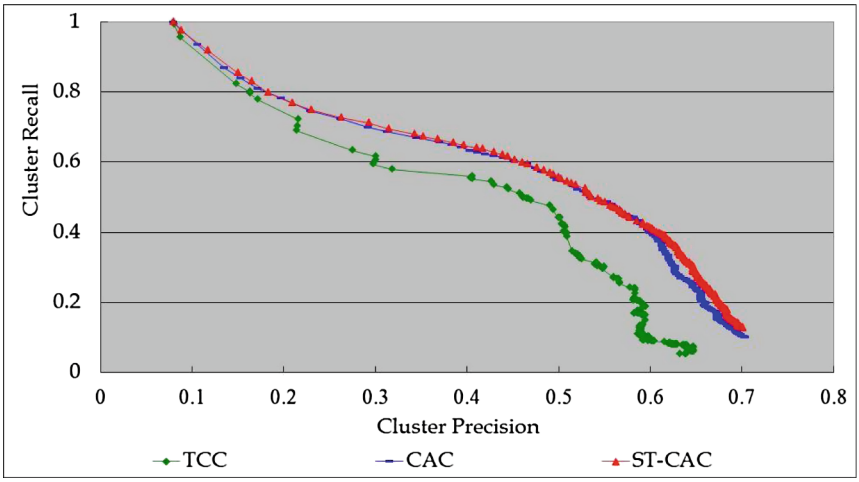

Fig. 2. Comparative evaluation results

5 Conclusion and Future Research Directions

Social media is nowadays an excellent source for gathering user intelligence to support various business intelligence applications. Motivated by the observation, this paper attempts to investigate the effectiveness of social tagging system (aka. folksonomy) in enhancing an important document management task, i.e., personalized document clustering. Specifically, we adopt the CAC technique proposed by Yang and Wei (2007) as our underlying algorithm and incorporate a leading social bookmarking site (i.e., Delicious) to design the ST-CAC technique which uses the folksonomy in Delicious to construct a statistical-based thesaurus for personalized document clustering. According to our empirical evaluation results, the ST-CAC and CAC techniques definitely have the ability to generate personalized document clusters than a traditional content-based approach. Moreover, the statistical-based thesaurus constructed from social media also slightly outperforms that generated from a general-purpose search engine.

Some ongoing and future research directions are briefly discussed as follows. First, Delicious, which is a social bookmarking service for webpages, is adopted as the social media for statistical-based thesaurus construction. Since our document corpus for evaluation purpose is collected from a scientific literature database, it is essential to evaluate the performance of an alternative social bookmarking service (i.e., CiteU-Like), which allows users to share citations to academic papers, on our proposed ST-CAC technique. Second, only the PMI measure is applied for statistical-based thesaurus construction. It should be interesting to implement and test empirically other measures for statistical-based thesaurus construction.

Acknowledgments. This work was supported by the National Science Council of the Republic of China under the grant NSC 100-2410-H-155-013-MY3 and the Ministry of Science and Technology of the Republic of China under the grant MOST 103-2410-H-155-027-MY3.

References

1. Aliakbary, S., Abolhassani, H., Rahmani, H., Nobakht, B.: Web page classification using social tags. In: International Conference on Computational Science and Engineering, pp. 588–593. IEEE Press, New York (2009)
2. Barreau, D.K.: Context as a factor in personal information management systems. J. Am. Soc. Inform. Sci. **46**, 327–339 (1995)
3. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G.: Social semantic query expansion. ACM Trans. Intell. Syst. Technol. **4**, 60 (2013)
4. Boley, D., Gini, M., Gross, R., Han, E., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., Moore, L.: Partitioning-based clustering for web document categorization. Decis. Support Syst. **27**, 329–341 (1999)
5. Brill, E.: A simple rule-based part of speech tagger. In: Third Conference on Applied Natural Language Processing, pp.152–155. Association for Computational Linguistics, Stroudsburg, PA (1992)

6. Cai, Y., Li, Q., Xie, H., Min, H.: Exploring personalized searches using tag-based user profiles and resource profiles in folksonomy. Neural Netw. **58**, 98–110 (2014)
7. Case, D.O.: Conceptual organization and retrieval of text by historians: the role of memory and metaphor. J. Am. Soc. Inform. Sci. **42**, 657–668 (1991)
8. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic grounding of tag relatedness in social bookmarking systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
9. Cutting, D., Karger, D., Pedersen, J., Tukey, J.: Scatter/gather: a cluster-based approach to browsing large document collections. In: 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 318–329. ACM Press, New York (1992)
10. Deogun, J., Raghavan, V.: User-oriented document clustering: a framework for learning in information retrieval. In: 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 157–163. ACM Press, New York (1986)
11. El-Hamdouchi, A., Willett, P.: Hierarchical document clustering using ward's method. In: 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 149–156. ACM Press, New York (1986)
12. eMarketer.: US social network users 2013: smartphone usage drives mobile-social growth (2007). https://www.emarketer.com/Coverage/SocialMedia.aspx
13. Facebook.: Facebook reports third quarter 2014 results (2014). http://investor.fb.com/releasedetail.cfm?ReleaseID=878726
14. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**, 265–323 (1999)
15. Kaplan, A.M., Haenlein, M.: Users of the world, unite! the challenges and opportunities of social media. Bus. Horiz. **53**, 59–68 (2010)
16. Kim, H., Lee, S.: A semi-supervised document clustering technique for information organization. In: 9th International Conference on Information and Knowledge Management, pp. 30–37. ACM Press, New York (2000)
17. Kwasnik, B.H.: The importance of factors that are not document attributes in the organization of personal documents. J. Doc. **47**, 389–398 (1991)
18. Lagus, K., Honkela, T., Kaski, S., Kohonen, T.: Self-organizing maps of document collections: a new approach to interactive exploration. In: 2nd International Conference on Knowledge Discovery and Data Mining, pp. 238–243. AAAI Press, Menlo Park (1996)
19. Larsen, B., Aone, C.: Fast and effective text mining using linear-time document clustering. In: 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 16–22. ACM Press, New York (1999)
20. Lin, C., Chen, H., Nunamaker, J.F.: Verifying the proximity and size hypothesis for self-organizing maps. J. Manage. Inform. Syst. **16**, 57–70 (1999–2000)
21. Milicevic, A.K., Nanopoulos, A., Ivanovic, M.: Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions. Artif. Intell. Rev. **33**, 187–209 (2010)
22. Movahedian, H., Khayyambashi, M.R.: Folksonomy-based user interest and disinterest profiling for improved recommendations: an ontological approach. J. Inform. Sci. **40**, 594–610 (2014)
23. Pantel, P., Lin, D.: Document clustering with committees. In: 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 199–206. ACM Press, New York (2002)
24. Pew Research Center: Social networking fact sheet (2014). http://www.pewinternet.org/fact-sheets/

25. Roussinov, D.G., Chen, H.: Document clustering for electronic meetings: an experimental comparison of two techniques. Decis. Support Syst. **27**, 67–79 (1999)
26. Rucker, J., Polanco, M.J.: Siteseer: personalized navigation for the web. Commun. ACM **40**, 73–75 (1997)
27. Suchanek, F.M., Vojnovic, M., Gunawardena, D.: Social tags: meaning and suggestions. In: 17th ACM Conference on Information and Knowledge Management, pp. 223–232. ACM Press, New York (2008)
28. Turney, P.D., Littman, M.L.: Measuring praise and criticism: inference of semantic orientation from association. ACM Trans. Inform. Syst. **21**, 315–346 (2013)
29. Vander Wal, T.: Folksonomy (2005). http://vanderwal.net/folksonomy.html
30. Voorhees, E.M.: Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. Inform. Process. Manage. **22**, 465–476 (1986)
31. Voutilainen, A.: Nptool: A detector of english noun phrases. In: Workshop on Very Large Corpora, pp. 48–57 (1993)
32. Wei, C., Chiang, R., Wu, C.: Accommodating individual categorization preferences: a personalized document clustering approach. J. Manage. Inform. Syst. **23**, 173–201 (2006)
33. Wei, C., Hu, P., Dong, Y.X.: Managing document categories in e-commerce environments: an evolution-based approach. Eur. J. Inform. Syst. **11**, 208–222 (2002)
34. Wu, C., Zhan, B.: Semantic relatedness in folksonomy. In: International Conference on New Trends in Information and Service Science, pp. 760–765. IEEE Press, New York (2009)
35. Yang, Y., Chute, C.G.: An example-based mapping method for text categorization and retrieval. ACM Trans. Inform. Syst. **12**, 252–277 (1994)
36. Yang, C.S., Wei, C.: Context-aware document-clustering technique. In: 11th Pacific Asia Conference on Information Systems (2007)
37. Yang, C.S., Chen, L.C.: Personalized recommendation in social media: a profile expansion approach. In: 18th Pacific Asia Conference on Information Systems (2014)