

# A Comparison of Different User-Similarity Measures as Basis for Research and Scientific Cooperation

Tamara Heck  
Heinrich-Heine-University  
Dept. for Information Science  
D-40225 Düsseldorf, Germany  
tamara.heck@uni-duesseldorf.de

New web-technologies can facilitate scientific and business work and research in many ways. The critical aspect is which structures and methods are to be used to elicit the best out of the existent resources? The information-overload is present in all-day life and sciences. Recommender systems try to solve this problem and have not only established in e-commerce, but also in the collaborative web, such as on Social-Bookmarking platforms [Heck/Peters 10]. In this paper, we analyze a database of records found on Bibsonomy, CiteULike and Connotea – three Social-Bookmarking Systems for scientific references – and explored the tripartite connection of users, documents and tags by three measurement methods. We concentrated on two research questions concerning the recommendation of similar users: 1) Are there differences when we apply different coefficients (Dice, Cosinus, Jaccard-Sneath)?, and 2) Are there differences when we apply shared documents or shared tags?

## 1. Related work

Different algorithms have been developed and analyzed to get the best performance for recommendation. We can decide between three methods: content-based and collaborative filtering recommender systems and hybrid forms of these two [Peters 09, Szomszor 07]. Social-Bookmarking Systems (SBS) mainly concentrate on the users' activities and therefore on collaborative filtering (CF), more precisely folksonomy-based CF: In a SBS, where user bookmark resources and tag them by keywords, the tripartite user-resource-tag connection can be used to recommend similar resources and also similar users. So far many algorithms concentrate on cocitation analysis [Van Eck/Waltman 08] and resource recommendation [e.g. Liang 08, Zanardi/Capra 08, Zhen/Li/Yeung 09], few on user recommendations [Luo et al. 08]. Comparisons of different similarity measures and algorithms can be found in [Cacheda et al. 11], [Egghe 09] and [Hamers et al. 89]. [Rorvig 99] concentrates on the visual exploration of measures based on different scaling methods. In this paper we also compare different coefficients, but in the second step compare the differences between user recommendation based on resources on the one hand and based on tags on the other.

## 2. Methods

In this paper we use the Dice, Cosinus and Jaccard-Sneath coefficient to measure the similarity between users of the SBS Bibsonomy, CiteULike and Connotea and analyze the different ranking results concerning the users' information needs.

According to [Van Eck/Waltman 08] a similarity measure (they used it for cocitations) should fulfill two conditions:

1. The similarity between two users should be maximal if the "profiles differ by at most a multiplicative constant" (p.1654).

2. There should be no similarity if the authors have nothing in common, i.e. any cocitations and in our case any bookmarks or tags.

$$S_{DiDj} = \frac{2g}{a + b} \quad S_{DiDj} = \frac{g}{\sqrt{a * b}} \quad S_{DiDj} = \frac{g}{a + b - g}$$

Dice Cosinus Jaccard-Sneath

Figure 1: Similarity coefficients used, where  $a$  is the number of single bookmarks or tags of User A,  $b$  the number of single bookmarks or tags of User B and  $g$  the number of common bookmarks or tags.

All three coefficients satisfy these conditions. Instead Pearson's correlation coefficient doesn't satisfy the conditions and shows some weaknesses (see e.g. Van Eck/Waltman 08), which was also discussed before by [Ahlgren/Jarvening/Rousseau 03]. In their paper they showed, among other things, that the Pearson correlation used for co-citation analysis has shortcomings when expanding the data sample, even if only zero-vector values are added (further discussion on the topic is also done by e.g. [Leydesdorff 05], [Leydesdorff/Vaughan 06], [Schneider/Borlund 07a] and [Schneider/Borlund 07b]).

Our database contains 13,762 bookmarks from CiteULike, Connotea and Bibsonomy, in our case scientific articles chosen from 45 physical journals. 10,498 of them are diverse articles, matched via DOI, title and UT-code. These bookmarks were tagged with 36,433 tags. We deleted the tags containing „%import%“, “%jabref%” and “%upload%” because these tags don't semantically describe the content of the bookmark they are generated to and the “file-import” tags were proposed automatically by CiteULike if a user imports his files. After this clearing we had 35,881 tags, which we revised further: lines and underlines were deleted, the plural forms replaced by singular forms and English words spelling with ‘s’ replaced by American spelling with ‘z’. This gave us 8,233 unique tags. We count 2,473 unique users who bookmarked the articles, 1,974 of them tagged their bookmarks.

We left out users who have only one bookmark because they would highly influence the results, i.e. user-pairs who have one bookmark in common and both only one bookmark at all, cause a similarity of 1. It would be important for a user-recommender system to set a threshold: either a user should have a minimum on bookmarks and/or he should have a minimum on similar bookmarks with another user, before this user is recommended to him. The last aspect can also be regulated by the user himself with the help of a slider, so the user can determine the amount of similar users who are recommended to him [see Knautz 10 for resource-recommendation]. CiteULike has a minimum of 20 resources a user must have in library before he gets resource recommendation. Leaving out all users with less than two bookmarks, we have 6,430 user-pairs who share at least one bookmark.

### 3. Results

#### 3.1 Differences between Coefficients

Analyzing our three coefficients we found correlation between user similarity based on resources and user similarity based on tags: In both scenarios Dice and Jaccard-Sneath gave similar results – the latter showing minor similarity value – cause the measure are quite similar (see [Egghe 10]). In contrast Cosinus gave different ranking results: It can be said, that Cosinus distinguishes between the allocation of the resources and tags between a user-pair [Hamers et al. 89]. The question is now which ranking of similar users serves best for the test-users' needs? If the test-user is searching for new resources, he might like a similar user who has bookmarked many resources.

In table 1 user “dchen”, who has 214 bookmarks, would be recommended user “caortiz” on rank 10 with a Dice similarity of 0.0185. Using Cosinus the similarity is 0.0967, with which “caortiz” would be

on rank three (Dice) or rank four (Cosinus). But for “dchen” the two bookmarks of “caortiz” possibly wouldn’t much help him for further research, contrary the 214 articles of “dchen” might be a good library for “caortiz”. If “dchen” would have less than 214 bookmarks, he would be more similar to “caortiz”. Concerning the research aspect for “caortiz” this would be a shortcoming for him.

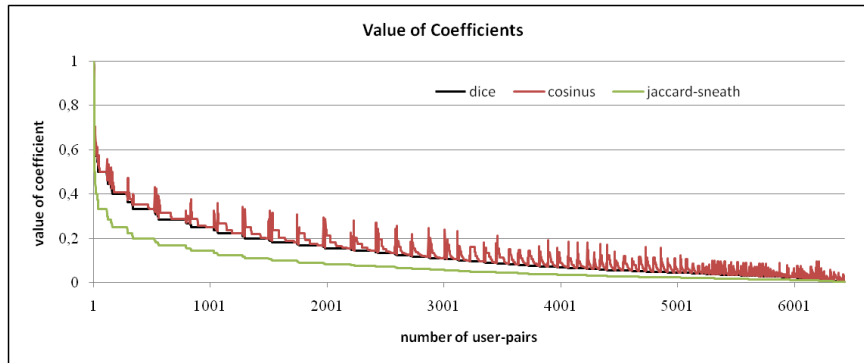


Figure 2: Value of coefficients by number of user-pairs: user-similarity based on bookmarks, source: Bibsonomy, CiteULike, Connotea.

common bm	bm dchen	bm user2	user1	user2	Dice bm	Cosinus bm
18	214	58	dchen	weeks	0.1324	0.1616
17	214	58	dchen	ghunter	0.125	0.1526
11	214	52	dchen	kdesmond	0.0827	0.1043
8	214	66	dchen	kkims	0.0571	0.0673
6	214	26	dchen	kedmond	0.05	0.0804
5	214	25	dchen	katiehumphry	0.0418	0.0684
4	214	15	dchen	tathabhatt	0.0349	0.0706
5	214	105	dchen	rodney	0.0313	0.0334
3	214	9	dchen	waitonhill	0.0269	0.0684
2	214	2	dchen	caortiz	0.0185	0.0967

Table 1: Similarity between user “dchen” and other SBS users based on common bookmarks (bm), ordered by Dice, source: Bibsonomy, CiteULike, Connotea.

### 3.2 Differences between tag- and resource-based similarities

Analyzing the user similarity based on tags, the coefficients provide similar results. The interesting aspect is the different ranking between similar users based on tags on the one hand and on resources on the other hand.

Apart from “ghunter” and “weeks, who both have the greatest similarity value either based on resources or on tags, there are users who wouldn’t be recommended to “dchen” if only common resources were taken into account. User similarity based on tags may have advantages over the one based on resources: There might be more and different similar users which wouldn’t be found over the resources, and tags may inform the user, in which context the other users read the resource. If “dchen” is searching for project partners, the tags can give him an impression of the article’s content the other users are interested in and therefore of the users’ research field.

common tags	tags dchen	tags user2	user1	user2	Dice tags	Cosinus tags
31	175	64	dchen	weeks	0.2594	0.2929
25	175	68	dchen	ghunter	0.2058	0.2292
20	175	29	dchen	kedmond	0.1961	0.2807
41	175	259	dchen	rodney	0.1889	0.1926
25	175	102	dchen	andreab	0.1805	0.1871
16	175	35	dchen	kkims	0.1524	0.2044
54	175	564	dchen	michaelbusmann	0.1461	0.1719
20	175	107	dchen	paulschlesinger	0.1418	0.1462
14	175	36	dchen	jeevanjyoti	0.1327	0.1764
23	175	176	dchen	bronckobuster	0.1311	0.1311

Table 2: Similarity between user “dchen” and other SBS users based on common tags, ordered by Dice, source: Bibsonomy, CiteULike, Connotea.

#### 4. Conclusion

The advantage of a user recommender system is that we could offer three recommendations: similar users for possible cooperation, relevant resources connected with users who “know” the research field (here the user himself can filter the results looking only at resources bookmarked by specific expert users), and tags which are assigned by specific expert users and lead to relevant resources. The user-recommendation therefore offers a further aspect which cannot be fulfilled with direct resource or tag recommendation. We also found out that there is a great difference between user recommendation based on shared resources and based on shared tags. Another aspect is the user’s needs: Is he searching for cooperation partners or for relevant resources? This is important for similarity measurement and the ranking of the results for a single user. Further research has to be done on this field. A new question is if there are features which could be integrated in user recommendation, for example information about which of the bookmarked papers the user has already read or which he dislikes. On CiteULike a user can now state if he has read an article or will read it; there is even a “Like it” button. These features might give further information about a user, which will be helpful to find cooperation partners and narrow the huge amount of relevant resources. Our further research tries to implement these aspects as well as to combine recommendation based on tags and on resources. Another investigation should also be the testing of different proximity and similarity measures. As several authors demonstrated (e.g. [Ahlgren/Jarvening/Rousseau 03], [Schneider/Borlund 07b]) the choice of a specific measure is often subjective. Different measures ask for different criteria and lead to different results, which require a detailed comparison.

#### 5. References

- [Ahlgren/Jarvening/Rousseau 03] Ahlgren, Per; Jarvening, Bo; Rousseau, Ronald (2003). Requirements for a Cocitation Similarity Measure, with Special Reference to Pearson’s Correlation Coefficient. In: Journal of the American Society for Information Science and Technology, 54, 6, 550-560.
- [Cacheda et al. 11] Cacheda, Fidel; Carneiro, Víctor; Fernández, Diego; Formoso, Vreixo (2011): Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. In: ACM Transactions on the Web, 5/1, article 2.

- [Egghe 10] Egghe, Leo (2010): Good Properties of Similarity Measures and Their Complementarity. In: Journal of the American Society for Information Science and Technology, 61/10, 2151–2160.
- [Hamers et al. 89] Hamers, Lieve; Hemeryck, Yves; Herweyers, Guido; Janssen, Marc (1989): Similarity Measures in Scientometric Research: The Jaccard Index Versus Salton's Cosine Formula. In: Information Processing & Management, 25/3, 315–318.
- [Heck/Peters 10]. Heck, Tamara; Peters, Isabella (2010). Expert Recommender Systems: Establishing Communities of Practice Based on Social Bookmarking Systems. In: Proceedings of I-Know 2010. 10<sup>th</sup> International Conference on Knowledge Management and Knowledge Technologies, 458-464.
- [Knautz 10] Knautz, Kathrin; Soubusta, Simone; Stock, Wolfgang G. (2010): Tag clusters as information retrieval interfaces. In: Proceedings of the 43th Annual Hawaii International Conference on System Sciences (HICSS-43), 10 pages.
- [Leydesdorff 05] Leydesdorff, Loet (2005). Similarity Measure, Author Cocitation Analysis, and Information Theory. In: Journal of the American Society for Information Science and Technology, 56, 7, 769-772.
- [Leydesdorff/Vaughan 06] Leydesdorff, Loet; Vaughan, Liwen (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. In: Journal of the American Society for Information Science and Technology, 57, 12, 1616–1628.
- [Liang et al. 08] Liang, Huizhi; Xu, Yue; Li, Yuefeng; Nayak, Richi (2008): Collaborative Filtering Recommender Systems Using Tag Information. In: ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2008 IEEE/WIC, 59–62.
- [Luo et al. 08] Luo, Heng; Niu, Changyong; Shen, Ruimin; Ullrich, Carsten (2008): A collaborative filtering framework based on both local user similarity and global user similarity. In: Machine Learning, 72/3, 231–245.
- [Peters 09] Peters, Isabella (2009): Folksonomies. Indexing and Retrieval in Web 2.0 (Knowledge and Information). De Gruyter, Saur: Berlin.
- [Rorvig 99] Rorvig, Mark (1999): Images of similarity: A Visual Exploration of Optimal Similarity Metrics and Scaling Properties of TREC Topic-Document Sets. In: Journal of the American Society for Information Science, 50/8, 639–651.
- [Schneider/Borlund 07a] Schneider, Jesper W.; Borlund, Pia (2007a): Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. In: Journal of the American Society for Information Science and Technology, 58, 11, 1586–1596.
- [Schneider/Borlund 07b] Schneider, Jesper W.; Borlund, Pia (2007b): Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the mantel and procrustes statistics. In: Journal of the American Society for Information Science and Technology, 58, 11, 1596–1609.
- [Szomszor et al. 07] Szomszor, M., Cattuto, C., Alani, H., O'Hara, K., Baldassarri, A., Loreto, V., Servedio, V. D. P. (2007): Folksonomies, the Semantic Web, and Movie Recommendation. In: 4th European Semantic Web Conference, Bridging the Gap between Semantic Web and Web 2.0, Innsbruck, Austria, 71-84.
- [Van Eck/Waltman 08] Van Eck, Nees Jan; Waltman, Ludo (2008): Appropriate Similarity Measures for Author Co-Citation Analysis. In: Journal of the American Society for Information Science and Technology, 59/10, 1653–1661.
- [Van Eck/Waltman 09] Van Eck, Nees Jan; Waltman, Ludo (2009): How to Normalize Cooccurrence Data? An Analysis of Some Well-Known Similarity Measures. In: Journal of the American Society for Information Science and Technology, 60/8, 1635–1651.
- [Zanardi/Capra 08] Zanardi, Valentina; Capra, Licia (2008): Social Ranking: Uncovering Relevant Content Using Tag-based Recommender Systems. In: Proceedings of the 2008 ACM Conference on Recommender Systems. ACM New York, NY, 51–58.
- [Zhen/Li/Yeung 09] Zhen, Yi; Li, Wu-Jun; Yeung, Dit-Yan (2009): TagiCoFi: tag informed collaborative filtering. In: Proceedings of the third ACM conference on Recommender systems, 69–76.