ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Improving accuracy by reducing the importance of hubs in nearest-neighbor recommendations

Clémentine Van Parijs & François Fouss

Université catholique de Louvain (UCL)
LSM & ICTEAM
Chaussée de Binche 151, 7000 Mons - Belgium

Abstract. A traditional approach for recommending items to persons consists of including a step of forming neighborhoods of users/items. This work focuses on such nearest-neighbor approaches and, more specifically, on a particular type of neighbors, the ones frequently appearing in the neighborhoods of users/items (i.e., very similar to many other users/items in the data set), referred to as *hubs* in the literature. The aim of this paper is to explore through experiments how the presence of hubs affects the accuracy of nearest-neighbor recommendations.

1 Introduction

Recommender systems try to provide people with recommendations of items they will appreciate, based on their past preferences, history of purchase, and demographic information (see, e.g., [1–3]). Three steps usually are common to recommender systems: (1) gather valuable information on the users and items, (2) determine patterns from the historical data, and (3) suggest items to people.

This paper relates to this second step of determining patterns from historical data. While content-based approaches (see, e.g., [2]) recommend items similar to the ones a user preferred in the past depending on the features of the items, collaborative approaches (see, e.g., [2]) recommend to a user items that people with similar tastes and preferences have liked (*user-based* recommandation) or items similar to the ones the considered user has preferred (*item-based* recommandation), depending, this time, on the links between items and users, and not on the features of items. These sets of similar users (items) are usually called the *nearest neighbors* of the user (of the items bought by the user). Our work focuses on a particular type of neighbors, the ones frequently appearing in the neighborhoods of users/items (i.e., very similar to many other people/items in the data set). As shown in other areas (see Section 2), such neighbors (referred to as *hubs* in the literature and therefore in this paper) play a very important role in nearest-neighbor processes.

This paper shows that applying nearest-neighbor methods in recommender systems also leads to the emergence of hubs and explore how the presence of hubs can affect the accuracy of nearest-neighbor recommendations. Our objective is therefore not to develop a state-of-the-art collaborative-recommendation method; rather, this paper aims at showing that the accuracy results obtained by state-of-the-art methods could be improved by reducing the importance of hubs. Section 2 mentions some work related to the presence of hubs and describes three intuitive approaches aiming at reducing the importance of hubs in nearest-neighbor recommendations. In Section 3, these three approaches are applied on two well-known data sets in the field of recommender systems, and results are shown and analyzed. Concluding remarks are discussed in Section 4.

2 Reducing the importance of hubs

The hubness phenomenon was first described in speech recognition [4], fingerprint identification [5] and music retrieval [6], further analyzed by Radovanović *et al.* in [7,8], Suzuki *et al.* in [9], Tomašev *et al.* in [10–12], and also studied in the field of collaborative filtering [13] where Nanopoulos *et al.* analyzed the hubness and the similarity-concentration phenomena.

As shown in these papers, neighbors frequently appearing in neighborhoods play an important role in nearest-neighbor processes in various areas. Our preliminary experiments on real data sets (see Section 3 for details about the data sets) confirmed that some users/items are included in an important number of neighborhoods. Following the intuition that neighbors common to lots of neighborhoods may bring less pertinent information for recommendations, since frequent, we formulate our research question as: “Could the accuracy of recommendations be influenced by neighbors belonging to many neighborhoods”?

Remember that this work relates to the second step common to all recommender systems, which aims at determining patterns from historical data. When based on neighborhoods, this second step is further divided into two substeps, the first one consists of computing similarities between users/items and the second one of forming neighborhoods (see, e.g., [14,15]). To answer our research question, three approaches (all lying between these two substeps since manipulating the similarities used for forming the neighborhoods) were developed aiming at mitigating the impact of the neighbors present in many neighborhoods.

The **first approach** (RMV) simply consists of *removing* from neighborhoods the *most frequent* neighbors. More precisely, k nearest neighbors of users/items are first identified (applying any method) and then the final neighborhoods are recomputed by removing the neighbors originally present in many neighborhoods. Notice that this technique therefore needs the tuning of a parameter, p, controlling the percentage of the initial neighbors deleted from neighborhoods.

Systematically ignore information that can be provided by some neighbors (such as in the first approach) can be considered as too radical. The idea of the **second approach** (NRM) is to first *normalize* the similarities (i.e., considering a user/item, each of its similarities - with other users/items - is first divided by the sum of its similarities with all the other users/items) in order to affect the similarity values associated to the neighbors and therefore the composition of the neighborhoods. Intuitively, the more a user/item is similar to many other users/items, the more its initial impact in neighborhoods will be reduced.

The **third approach** (RNK) relies on *ranks* to select the k nearest neighbors of a user/item. For a considered user/item, each similarity - i.e., with another user/item - is simply replaced by its rank among the similarities between the considered user/item and all the other users/items (i.e., a 1 replaces the highest similarity of the considered user, a 2 the second highest, etc.). The neighborhoods are now formed using these ranks rather than initial similarities. Applying this procedure limits the number of times a neighbor (originally present in many neighborhoods) appears in the new ones while every user/item is somehow forced to be present in at least some neighborhoods. The intuition is that every user/item should be the neighbor of at least some others, which is quite natural except for, rare, very atypical, users/items.

Notice that the second and third approaches are free of parameters.

3 Experiments

Two different data sets, well-known in recommender systems, were used for experiments in this work. The MovieLens (ML) data set contains 100.000 ratings of 943 persons about 1.682 movies and the BookCrossing (BC) data set contains 109.374 ratings of 1.028 persons about 2.222 books. Note that, for all the experiments, the numerical value of the ratings provided by the users are not taken into account, but only the fact that a user watched a movie / bought a book.

The criteria used in this work for assessing the accuracy of the approaches is the recall score, averaged on all the users, quantifying for each user the proportion of the top N recommended items that should be recommended to the user, according to historical data. The recall scores should be as high as possible (i.e., close to 100%) for good performance and are computed through a classical double cross-validation prodecure: An internal cross-validation is used for tuning the parameters (i.e., the number of neighbors k $(= 10, 20, ..., 100)$ and the percentage of removed users in the first approach $p = (5\%, 10\%, ..., 40\%)$), and an external cross-validation for assessing the approach when the values of the parameters are fixed to the ones providing the best accuracy results in internal cross-validation. Details of the computation of the recommendations and recall scores (recall 10 and recall 20 are reported in this paper) as well as details on the applied double cross-validation procedure can be found in [14].

To evaluate if the three approaches improve the accuracy of recommendations, the so-called *Bin* method was chosen as a reference, for its simplicity and for providing very competitive results when applied on the MovieLens and BookCrossing data sets (see [14, 15]). Applied in the user-based method (the item-based method is similar and does not require further explanations), each user i is characterized by a binary vector (whose dimension is the total number of items) encoding the items he bought. Similarities between pairs of vectors (and therefore pairs of users) are then computed. In [15], systematic comparisons between eight such measures (listed in [16], p. 674) were performed. The best recall scores were obtained with the measure "ratio of 1-1 matches to mismatches with 0-0 matches excluded", therefore also used in these experiments.

Results. Table 1 shows the recall scores obtained on both data sets and both methods (user-based, item-based). For each case, the recall obtained in the reference situation (REF) and the ones obtained by applying the three approaches modifying the composition of the neighborhoods (i.e., the removal of hubs (RMV), the normalization technique (NRM), and the ranks technique (RNK)) are shown. We used a paired t-test to determine whether there is a significant difference (with a p-value smaller than 0.01) between the recalls; the results that are significantly better than REF are in bold. The improvements of the recalls (in percentage) are also shown, in the "Variation" rows.

The variation of the recalls when removing some neighbors present in many neighborhoods are rather low (close to 1%), and are positive in the user-based method and negative in the item-based method. Nevertheless, it shows that disputing the presence of some neighbors may be beneficial. Particularly, in the case of the ML data set, it appeared that we could remove up to 35% of hubs and still have improvements with respect to the initial situation.

The NRM approach leads to better accuracy results for both data sets and both methods. Improvements, generally significant, are important: between 4.00

	MovieLens dataset				**BookCrossing dataset**			
	User-based method							
in %	REF	RMV	NRM	RNK	REF	RMV	NRM	RNK
Recall 10	24.66	25.05	**25.74**	**25.74**	7.23	7.30	7.51	**7.57**
Variation	/	1.61	4.40	4.40	/	1.04	4.00	4.76
Recall 20	35.88	**36.39**	**36.66**	**37.22**	11.19	11.35	**11.59**	11.63
Variation	/	1.41	2.16	3.72	/	1.41	3.60	3.94
	Item-based method							
Recall 10	22.05	22.05	**24.43**	**25.32**	9.41	9.29	**9.89**	**10.05**
Variation	/	0.01	10.79	14.82	/	−1.34	5.04	6.79
Recall 20	34.56	34.54	**35.95**	**37.59**	13.96	13.76	**14.55**	**14.82**
Variation	/	−0.08	4.02	8.76	/	−1.49	4.20	6.17

Table 1: Recall scores obtained on both data sets, both methods for the reference (REF) and our three approaches (RMV, NRM, and RNK).

MovieLens dataset			
	REF vs NRM	REF vs RNK	NRM vs RNK
User-based method	60.68	54.46	76.28
Item-based method	48.90	49.00	57.40
BookCrossing dataset			
	REF vs NRM	REF vs RNK	NRM vs RNK
User-based method	51.90	57.12	81.84
Item-based method	51.23	62.33	70.33

Table 2: Percentage of common neighbors for both data sets and both methods.

and 10.79% for the recall 10, and from 2.16 to 4.20% for the recall 20. These are particularly good for the item-based method.

Improvements for the RNK approach are the best ones, from 4.40 to 14.82% for the recall 10 and from 3.72 to 8.76% for the recall 20. The pattern is a bit the same as for the normalization: improvements are important in each case, but particularly for the item-based method. A consequence is that, for the ML data set, recall scores obtained with the item-based method compete with those obtained with the user-based method.

Further analysis. These experiments suggest that our two last approaches - NRM and RNK - may substantially improve the accuracy of recommendations. This gives rise to the conclusion that trying to reduce the impact of hubs in neighborhoods can be a good option. It is therefore interesting to find out to what extend the composition of neighborhoods has evolved with the two last approaches, in comparison with the reference situation. The **percentage of common neighbors** when comparing the neighborhoods of a user computed by various approaches (REF, NRM, RNK) is first computed. Table 2 reports the average of these percentages on each user/item and on each run of the cross-validation. It appears that the NRM and RNK approaches both modify the initial composition of neighborhoods by around $40 - 50\%$. The aim of the last two approaches being the same, it is quite logical to observe that they present a high percentage of common neighbors (between 57.40 and 81.81%).

If neighborhoods are more balanced, another question is about their **reciprocity** which can be, for example, quantified by the number of users/items that are present in neighborhoods of their neighbors. For that, we computed the average number of times an element is present in the neighborhoods of his neighbors. Results, averaged over the ten runs of the cross-validation and expressed as a percentage, are shown in Table 3. A similar trend is observed in most of the cases (exception is NRM for the item-based method on the BC data set): neighborhoods are more reciprocal with NRM than in REF, and still a bit

MovieLens dataset			
	REF	NRM	RNK
User-based method	48.37	66.17	70.61
Item-based method	44.15	44.35	56.65
BookCrossing dataset			
	REF	NRM	RNK
User-based method	50.05	62.98	75.52
Item-based method	56.60	51.36	76.37

Table 3: Reciprocity in neighborhoods for both data sets and both methods.

MovieLens dataset						
	User-based method			Item-based method		
	REF	NRM	RNK	REF	NRM	RNK
Novelty 10	288.86	258.52	277.51	202.62	253.54	235.01
Novelty 20	245.69	217.50	234.91	190.11	233.83	207.73
BookCrossing dataset						
	User-based method			Item-based method		
	REF	NRM	RNK	REF	NRM	RNK
Novelty 10	174.66	129.95	151.46	102.37	121.23	105.59
Novelty 20	161.28	122.48	139.98	99.25	114.26	96.99

Table 4: Novelty for both data sets and both methods.

more with RNK.

Till now, our analysis focused on the accuracy of recommendations. By analyzing the new neighborhoods, we wondered whether recommendations could be more surprising with NRM or RNK. Indeed, reducing the impact of hubs could intuitively lead to reduce the recommendation of items with a high frequency. To this end, a **novelty** score (the average of the median frequency of 10 or 20 recommended items) was computed and the results are shown in Table 4 (this measure should be as low as possible for a good performance in terms of novelty). In REF, the item-based method always provides better results than the user-based method. When looking at NRM and RNK, results are essentially better for the user-based method.

To summarize, using NRM or RNK in the user-based method improves the recall scores and provides more surprising recommendations while important accuracy improvements were obtained with the item-based method, but without leading to more surprising recommendations.

4 Conclusion

The hubness phenomenon inherent to nearest-neighbor methods has been highlighted in many different areas such as, e.g., speech recognition, clustering, classification, or even collaborative filtering where [13] questioned the meaning and the representativeness of discovered nearest neighbors. This work goes one step further by experimentally showing that the presence of hubs can affect the accuracy of nearest-neighbor recommendations and suggesting three intuitive approaches aiming at mitigating the impact of the neighbors present in many neighborhoods. Results showed that two approaches (i.e., NRM and RNK) significantly improve the accuracy of the recommendations, lead to more symmetric neighborhoods and also, in different cases, to more surprising recommendations. Further work will include the strengthening of our analysis by including other performance metrics (such as, e.g., precision and F-score) and the investigation of this hubness phenomenon on the behavior of other nearest-neighbors based methods (such as, e.g., graph-based methods).

References

[1] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems: An introduction.* Cambridge University Press, 2011.

[2] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou. Recommender systems. *Physics Reports*, 519:1–49, 2012.

[3] F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor. *Recommender Systems Handbook.* Springer, 2011.

[4] G. Doddington, W. Ligget, A. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *Proceedings of the 5th International Conference on Spoken Language Processing*, 1998.

[5] A. Hicklin, C. Watson, and B. Ulery. The myth of goats: How many people have fingerprints that are hard to match? *Internal Report 7271, National Institute of Standards and Technology*, 2005.

[6] J.-J. Aucouturier and F. Pachet. A scale-free distribution of false positives for a large class of audio similarity measures. *Pattern Recognition*, 41(1):272–284, 2007.

[7] M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

[8] M. Radovanović, A. Nanopoulos, and M. Ivanović. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd Annual International Conference on Research and Development in Information Retrieval*, pages 186–193, 2010.

[9] I. Suzuki, K. Hara, M. Shimbo, Y. Matsumoto, and M. Saerens. Investigating the effectiveness of Laplacian-based kernels in hub reduction. *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 1112–1118, 2012.

[10] N. Tomašev, R. Brehar, D. Mladenič, and S. Nedevschi. The influence of hubness on nearest-neighbor methods in object recognition. *Proceedings of the IEEE International Conference on Intelligent Computer Communication and Processing*, pages 367–374, 2011.

[11] N. Tomašev and D. Mladenič. Nearest neighbor voting in high dimensional data: learning from past occurences. *Computer Science and Information Systems*, 9(2):691–712, 2012.

[12] N. Tomašev, M. Radovanovič, D. Mladenič, and M. Ivanovič. The role of hubness in clustering high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 99, 2013.

[13] A. Nanopoulos, M. Radovanovič, and M. Ivanovič. How does high dimensionality affect collaborative filtering? *Proceedings of the 3rd ACM Conference on Recommender Systems*, pages 293–296, 2009.

[14] F. Fouss, K. Francoisse, L. Yen, A. Pirotte, and M. Saerens. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural Networks*, 31:53–72, 2012.

[15] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, 2007.

[16] R. Johnson and D. Wichern. *Applied multivariate statistical analysis, 5th Ed.* Prentice Hall, 2002.