

Improvement of similarity-diversity trade-off in recommender systems based on a facility location model

ANTIOPI PANTELI*[§] and BASILIS BOUTSINAS†[§]

[§] *Dept. of Business Administration, MIS & Business Intelligence Lab, University of Patras,
Rio Patras, 26500, Greece*

There is a growing interest in the offering of novel alternative choices to users of recommender systems. These recommendations should match the target query while at the same time they should be diverse with each other in order to provide useful alternatives to the user, i.e. novel recommendations. In this paper, the problem of extracting novel recommendations, under the similarity-diversity trade-off, is modeled as a facility location problem. We formulate this trade-off as a multiple p -median problem solved by using biclustering. The results from tests in the benchmark Travel Case Base were satisfactory when compared to well-known recommender techniques, in terms of both similarity and diversity. Moreover, the experimental tests have shown that the proposed method is flexible enough, since a parameter of the adopted facility location model constitutes a regulator for the trade-off between similarity and diversity.

Keywords: Recommender Systems; similarity-diversity trade-off; facility location model; biclustering.

1. Introduction

The development of the Internet has resulted in an overload of data and users often find it difficult to extract information that best corresponds to their preferences or needs. Thus, recommender systems became part of life since they can manage and process the available information in order to filter the redundant part and extract useful knowledge (Perugini *et al.*, 2004). Their applications can be found in different aspects of everyday life such as health (Trang Tran *et al.*, 2017), music (Chen & Chen, 2005), movies (Golbeck, 2006), travel (Fesenmaier *et al.*, 2003; Shih *et al.*, 2011), news (H. J. Lee & Park, 2007), e-learning (Bobadilla *et al.*, 2009) etc.

There are some major types of recommender systems that have been studied in the literature, such as the content-based systems (CBS), the collaborative filtering approaches (CF), the knowledge-based (KBS) systems, as well as hybrid approaches. Case-based recommender systems that are studied in this paper constitute a sub-class of knowledge-based systems (KBS). In a case-based recommender system, the user inserts a query and the system, based on a similarity measure, tries to match the query to the already recorded cases that exist in the Case Base.

* Corresponding author. Email: antiopi.panteli@gmail.com

† Email: yutsinas@upatras.gr

It is a current trend that the recommendations provided to the user have to be quite similar to the inserted query, but also they have to be different to each other in order to offer diverse options to the user (Hurley & Zhang, 2011; Smyth & McClave, 2001; Zhou *et al.*, 2010; Ziegler *et al.*, 2005), since users may get bored after receiving many recommended objects under the same topic. However, the most common problem in the study of recommender systems, is the trade-off between similarity and diversity.

In this work, we apply a facility location model to improve the similarity-diversity trade-off in recommender systems. Our work can broaden the perspectives of the interaction and combination of different scientific fields in order to achieve the best possible results. We evaluate the proposed approach by using a benchmark recommender system that offers alternative choices in order to propose diverse travel recommendation plans to users, by selecting a specific number of cases from the Case Base that match the user's requirements and preferences expressed in the target query.

An important characteristic of the proposed approach is that a specific parameter of the facility location model constitutes a regulator for the trade-off between similarity and diversity of the recommendation set.

2. Background

Most works in the literature related to diversity enhancement in recommendation lists refer to collaborative filtering approaches rather than to case-based approaches. Such approaches can be found in the papers of Abbassi *et al.* (2013), Adomavicius & Kwon (2012), Aytakin & Karakaya (2014), Boim *et al.* (2011), Choi & Han (2010), Zhang & Hurley (2008) and Ziegler *et al.* (2005). The common problem in recommendation strategies is the trade-off between similarity and diversity. The trade-off between diversity and similarity for case-based recommender systems has been studied much less than for collaborative filtering approaches. This is one of the reasons that motivated our research in this topic. However, it should be mentioned that there are some research works related to this trade-off in case-based recommender systems (Bradley & Smyth (2001); Hurley & Zhang, 2011; McSherry, 2002; Smyth & McClave, 2001).

2.1 The *P*-median Problem (PMP)

In this subsection, we give a brief overview of a well-known facility location problem, the *p*-median problem, which was introduced by Hakimi (1964). It consists of locating *p* facilities that serve a given set of demand points such that the total distance between demand points and facilities is minimized.

Since the PMP is NP-hard (Kariv & Hakimi, 1979), exact methods can be used efficiently for smaller instances. However, for larger instances, e.g. including 1000 nodes, exact methods become inefficient and heuristic approaches have been proposed

by providing solutions slightly inferior to the optimal ones in much less computational time than the exact methods. The first formulation of the problem as a zero-one programming problem can be found in the paper of ReVelle & Swain (1970).

Since uncertainty is inherent in most real life applications, there is often a possibility that each facility may fail and may not be able to satisfy the given demand due to weather, labor actions, electricity problems and other factors. In Panteli *et al.* (2019), an extension of the original p -median problem, the multiple p -median problem (MPMP), was solved. This problem focuses on the possibility of service of a demand point more than one times, expressed by a parameter (mc).

The objective of the problem is to select p facilities represented by columns in order to serve each demand point at least mc times, each time from a different facility. The facilities have to be selected in such a way that the total distance between demand points and facilities is the minimum possible. The decision variables of the problems have to be defined at first:

$$x_{ij} = \begin{cases} 1, & \text{if demand point } j \text{ is served by a facility located at site } i \\ 0, & \text{otherwise} \end{cases}$$

$$y_i = \begin{cases} 1, & \text{if a facility is located at candidate site } i \\ 0, & \text{otherwise} \end{cases}$$

Thus, the formulation of the MPMP is the following:

$$\begin{aligned} & \text{Minimize } \sum_{i \in I} \sum_{j \in J} d_{ij} x_{ij} \\ & \text{subject to } \sum_{i \in I} x_{ij} \geq mc, \forall j \in J \\ & \sum_{i \in I} y_i = p \\ & x_{ij} - y_i \leq 0, \forall i \in I; j \in J \\ & y_i \in \{0, 1\}, \forall i \in I \\ & x_{ij} \in \{0, 1\}, \forall i \in I; j \in J \end{aligned}$$

The adopted heuristic algorithm for the MPMP takes as input a Boolean data matrix $D(R, C)$. Cell values of $D(R, C)$ represent the weight (e.g. distance) between specific demand points (represented by rows) and facilities (represented by columns).

In this paper, we broaden the perspectives of applications of the MPMP since the backup option does not represent only the notion of facility failure in covering the demand but also it represents the notion of multiple options. Thus, it represents the alternative choices which are recommended to the user by a recommender system. These alternatives have to match the target query's characteristics and at the same

time have to be diverse with each other, in order to constitute useful alternatives to the user, i.e. novel recommendations.

The multiple p-median facility location problem can be applied to recommender systems in order to offer multiple case recommendations to the user, exploiting the parameter (mc) which refers to the number of times that a demand point should be served. The recommendation set presents higher or lower diversity and similarity ratios based on the values of the parameter mc .

3. Experimental Results

In order to evaluate the proposed methodology, we used the standard benchmark case library for recommender systems, the Travel Case Base which contains categorical and numerical data. This case base has been appropriately transformed in order to be used by the adopted method for solving the MPMP model.

Table 1 presents the experimental results of the application of the multiple p-median model to recommender systems. The similarity and diversity ratio are calculated as in Smyth & McClave (2001).

The Travel Case Base contains 1024 cases but no real user queries. Thus, we used for our experiments the leave-one-out method to obtain queries.

TABLE 1 *Experimental Results*

mc	p	Diversity	Similarity	Similarity to Diversity Ratio
3	5	0.33783	0.73169	2.166
4	5	0.31407	0.73295	2.334
5	5	0.25923	0.73321	2.828
3	10	0.36108	0.70065	1.940
5	10	0.35329	0.70068	1.983
7	10	0.33663	0.69950	2.078
10	10	0.30170	0.69990	2.320
5	15	0.37412	0.67896	1.815
10	15	0.35160	0.66820	1.900
15	15	0.34350	0.66370	1.932
5	20	0.42599	0.63301	1.486
10	20	0.40553	0.63494	1.566
20	20	0.38848	0.62245	1.602

5.1. Comparison with other methods

Table 2 presents the results of the proposed method with respect to the alternative diversity-preserving techniques in literature.

TABLE 2 *Effectiveness of the proposed method with respect to alternative techniques in literature*

		Standard	Bounded Greedy	Random	Greedy	Proposed Method (Leave-One-Out)	Proposed Method (400 Queries)
		S	BG	R	G	PM	PM
$p=5$	Similarity	0.78	0.753	0.748	0.7	0.73169	0.69
	Diversity	0.289	0.375	0.326	0.458	0.33783	0.41

4. Conclusions

The application of a facility location model to recommender systems provides effectiveness and flexibility in terms of similarity of the recommending set when compared to the target query and diversity between the recommended cases with each other. The mc parameter constitutes a regulator for the trade-off between similarity and diversity of the recommendation set. Moreover, the proposed approach performs well when compared to the alternative diversity preserving techniques, while at the same time it exhibits very fast computation times with respect to these approaches.

Acknowledgement

The authors wish to thank Aristotelis Kompothrekas, PhD candidate in Patras University, for his valuable contribution in performing the experimental tests.

REFERENCES

- ABBASSI, Z., MIRROKNI, V. S., & THAKUR, M. (2013). *Diversity maximization under matroid constraints*. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.
- ADOMAVICIUS, G., & KWON, Y. (2012). Improving aggregate recommendation diversity using ranking-based techniques. *IEEE transactions on knowledge and data engineering*, 24(5), 896-911.
- AGRAWAL, R., GOLLAPUDI, S., HALVERSON, A., & IEONG, S. (2009). *Diversifying search results*. In: Proceedings of the second ACM international conference on web search and data mining.
- ALP, O., ERKUT, E., & DREZNER, Z. (2003). An efficient genetic algorithm for the p-median problem. *Annals of Operations Research*, 122(1), 21-42. doi: 10.1023/A:1026130003508
- AVELLA, P., BOCCIA, M., SALERNO, S., & VASILYEV, I. (2012). An aggregation heuristic for large scale p-median problem. *Computers & Operations Research*, 39(7), 1625-1632. doi: 10.1016/j.cor.2011.09.016
- AYTEKIN, T., & KARAKAYA, M. Ö. (2014). Clustering-based diversity improvement in top-N recommendation. *Journal of Intelligent Information Systems*, 42(1), 1-18. doi: 10.1007/s10844-013-0252-9
- BEASLEY, J. E. (1985). A note on solving large p-median problems. *European Journal of Operational Research*, 21(2), 270-273. doi: 10.1016/0377-2217(85)90040-2
- BEN-DOR, A., CHOR, B., KARP, R., & YAKHINI, Z. (2003). Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of computational biology*, 10(3-4), 373-384. doi: 10.1089/10665270360688075
- BOBADILLA, J., SERRADILLA, F., & HERNANDO, A. (2009). Collaborative filtering adapted to recommender systems of e-learning. *Knowledge-Based Systems*, 22(4), 261-265.
- BOIM, R., MILO, T., & NOVGORODOV, S. (2011). *Diversification and refinement in collaborative filtering recommender*. In: Proceedings of the 20th ACM international conference on Information and knowledge management.
- BOUTSINAS, B. (2013a). Machine-part cell formation using biclustering. *European Journal of Operational Research*, 230(3), 563-572. doi: 10.1016/j.ejor.2013.05.007
- BOUTSINAS, B. (2013b). A new biclustering algorithm based on association rule mining. *International Journal on Artificial Intelligence Tools*, 22(03), 1350017. doi: 10.1016/j.ejor.2013.05.007
- BOUTSINAS, B., & PAPASTERGIOU, T. (2008). On clustering tree structured data with categorical nature. *Pattern recognition*, 41(12), 3613-3623.
- BOUTSINAS, B., SIOTOS, C., & GEROLIMATOS, A. (2008). Distributed mining of association rules based on reducing the support threshold. *International Journal on Artificial Intelligence Tools*, 17(06), 1109-1129. doi: 10.1142/S0218213008004321
- BOZKAYA, B., ZHANG, J., & ERKUT, E. (2002). An efficient genetic algorithm for the p-median problem. *Facility location: Applications and theory*, 179-205.
- BRADLEY, K., & SMYTH, B. (2001). Improving recommendation diversity. In: Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland.
- BUSYGIN, S., PROKOPYEV, O., & PARDALOS, P. M. (2008). Biclustering in data mining. *Computers & Operations Research*, 35(9), 2964-2987. doi: 10.1016/j.cor.2007.01.005
- CHEN, H.-C., & CHEN, A. L. (2005). A music recommendation system based on music and user grouping. *Journal of Intelligent Information Systems*, 24(2), 113-132. doi: 10.1007/s10844-005-0319-3
- CHENG, Y., & CHURCH, G. M. (2000). *Biclustering of expression data*. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology.
- CHOI, S.-M., & HAN, Y.-S. (2010). *A content recommendation system based on category correlations*. In: Computing in the Global Information Technology (ICCGI), 2010 Fifth International Multi-Conference on.
- CLARKE, C. L., et al. (2008). *Novelty and diversity in information retrieval evaluation*. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.

- DAS SARMA, A., GOLLAPUDI, S., & IEONG, S. (2008). *Bypass rates: reducing query abandonment using negative inferences*. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.
- DENSHAM, P. J., & RUSHTON, G. (1992). A more efficient heuristic for solving large p-median problems. *Papers in regional science*, 71(3), 307-329. doi: 10.1111/j.1435-5597.1992.tb01849.x
- DEODHAR, M., & GHOSH, J. (2007). *A framework for simultaneous co-clustering and learning from complex data*. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.
- DHILLON, I. S. (2001). *Co-clustering documents and words using bipartite spectral graph partitioning*. In: Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining.
- DREZNER, Z., & HAMACHER, H. W. (2001). *Facility location: applications and theory*: Springer Science & Business Media.
- FESENMAIER, D. R., *et al.* (2003). *DIETORECS: Travel advisory for multiple decision styles*.
- FITZSIMMONS, J. A., & ALLEN, L. A. (1983). A warehouse location model helps Texas comptroller select out-of-state audit offices. *Interfaces*, 13(5), 40-46. doi: 10.1287/inte.13.5.40
- FLEDER, D., & HOSANAGAR, K. (2009). Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science*, 55(5), 697-712.
- GARY, M. R., & JOHNSON, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-completeness*: WH Freeman and Company, New York.
- GOLBECK, J. (2006). *Generating predictive movie recommendations from trust in social networks*. In: International Conference on Trust Management.
- GUNAWARDANA, A., & SHANI, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10(Dec), 2935-2962.
- HAKIMI, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, 12(3), 450-459. doi: 10.1287/opre.12.3.450
- HAN, J., PEI, J., YIN, Y., & MAO, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1), 53-87. doi: 10.1023/B:DAMI.0000005258.31418.83
- HANJOUL, P., & PEETERS, D. (1985). A comparison of two dual-based procedures for solving the p-median problem. *European Journal of Operational Research*, 20(3), 387-396. doi: 10.1016/0377-2217(85)90012-8
- HANSEN, P., & MLADENović, N. (1997). Variable neighborhood search for the p-median. *Location Science*, 5(4), 207-226. doi: 10.1016/S0966-8349(98)00030-8
- HARTIGAN, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337), 123-129. doi: 10.1080/01621459.1972.10481214
- HERLOCKER, J. L., KONSTAN, J. A., TERVEEN, L. G., & RIEDL, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1), 5-53. doi: 10.1145/963770.963772
- HUANG, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3), 283-304. doi: 10.1023/A:1009769707641
- HURLEY, N., & ZHANG, M. (2011). Novelty and diversity in top-n recommendation--analysis and evaluation. *ACM Transactions on Internet Technology (TOIT)*, 10(4), 14. doi: 10.1145/1944339.1944341
- KARIV, O., & HAKIMI, S. L. (1979). An algorithmic approach to network location problems. II: The p-medians. *SIAM Journal on Applied Mathematics*, 37(3), 539-560. doi: 10.1137/0137041
- LAZZERONI, L., & OWEN, A. (2002). Plaid models for gene expression data. *Statistica sinica*, 61-86.
- LEE, H. J., & PARK, S. J. (2007). MONERS: A news recommender for the mobile web. *Expert Systems with Applications*, 32(1), 143-150.
- LIU, J., & WANG, W. (2003). *OP-Cluster: Clustering by Tendency in High Dimensional Space*. In: Proceedings of the Third IEEE International Conference on Data Mining.
- LORENZI, F., RICCI, F., TOSTES, R., & BRASIL, R. (2005). Case-based recommender systems: A unifying view. *Lecture notes in computer science*, 3169, 89.
- MADEIRA, S. C., & OLIVEIRA, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1), 24-45. doi: 10.1109/TCBB.2004.2
- MARANZANA, F. (1964). On the location of supply points to minimize transport costs. *OR*, 261-270. doi: 10.2307/3007214

- MCSHERRY, D. (2002). Diversity-Conscious Retrieval. In S. Craw & A. Preece (Eds.), *Advances in Case-Based Reasoning: 6th European Conference, ECCBR 2002 Aberdeen, Scotland, UK, September 4–7, 2002 Proceedings* (pp. 219-233). Berlin, Heidelberg: Springer Berlin Heidelberg.
- MEGIDDO, N. (1986). *On the complexity of linear programming*: IBM Thomas J. Watson Research Division.
- MITROPOULOS, P., MITROPOULOS, I., & GIANNIKOS, I. (2013). Combining DEA with location analysis for the effective consolidation of services in the health sector. *Computers & Operations Research*, 40(9), 2241-2250.
- MITROPOULOS, P., MITROPOULOS, I., GIANNIKOS, I., & SISOURLAS, A. (2006). A biobjective model for the locational planning of hospitals and health centers. *Health Care Management Science*, 9(2), 171-179.
- MLADENović, N., BRIMBERG, J., HANSEN, P., & MORENO-PÉREZ, J. A. (2007). The p-median problem: A survey of metaheuristic approaches. *European Journal of Operational Research*, 179(3), 927-939. doi: 10.1016/j.ejor.2005.05.034
- MUCHERINO, A., & URTUBIA, A. (2010). *Consistent Biclustering and Applications to Agriculture*. In: Industrial Conference on Data Mining-Workshops.
- PANTELI, A., BOUTSINAS, B., & GIANNIKOS, I. (2018). On solving the multiple p-median problem in large scale data based on biclustering. *Operational Research*, to appear.
- PEKER, S., & KOCYIGIT, A. (2016). *An Adjusted Recommendation List Size Approach for Users' Multiple Item Preferences*. In: International Conference on Artificial Intelligence: Methodology, Systems, and Applications.
- PERUGINI, S., GONÇALVES, M. A., & FOX, E. A. (2004). Recommender Systems Research: A Connection-Centric Survey. *Journal of Intelligent Information Systems*, 23(2), 107-143. doi: 10.1023/b:jiis.0000039532.05533.99
- PU, P., et al. (2011). Usability guidelines for product recommenders based on example critiquing research *Recommender systems handbook* (pp. 511-545): Springer.
- REVELLE, C. S., & SWAIN, R. W. (1970). Central facilities location. *Geographical Analysis*, 2(1), 30-42. doi: 10.1111/j.1538-4632.1970.tb00142.x
- RICCI, F., ROKACH, L., & SHAPIRA, B. (2011). Introduction to recommender systems handbook *Recommender systems handbook* (pp. 1-35): Springer.
- ROSENWEIN, M. B. (1994). Discrete location theory, edited by P. B. Mirchandani and R. L. Francis, John Wiley & Sons, New York, 1990, 555 pp. *Networks*, 24(2), 124-125. doi: 10.1002/net.3230240212
- SARWAR, B., KARYPIS, G., KONSTAN, J., & RIEDL, J. (2000a). *Analysis of recommendation algorithms for e-commerce*. In: Proceedings of the 2nd ACM conference on Electronic commerce.
- SARWAR, B., KARYPIS, G., KONSTAN, J., & RIEDL, J. (2000b). Application of dimensionality reduction in recommender system-a case study: Minnesota Univ Minneapolis Dept of Computer Science.
- SCOTT, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3), 605-610.
- SHIH, D.-H., YEN, D. C., LIN, H.-C., & SHIH, M.-H. (2011). An implementation and evaluation of recommender systems for traveling abroad. *Expert Systems with Applications*, 38(12), 15344-15355.
- SMYTH, B., & MCCLAVE, P. (2001). *Similarity vs. Diversity*. In: Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development.
- STURGES, H. A. (1926). The choice of a class interval. *Journal of the american statistical association*, 21(153), 65-66.
- TANAY, A., SHARAN, R., & SHAMIR, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(suppl 1), S136-S144. doi: 10.1093/bioinformatics/18.suppl_1.S136
- TEITZ, M. B., & BART, P. (1968). Heuristic methods for estimating the generalized vertex median of a weighted graph. *Operations Research*, 16(5), 955-961. doi: 10.1287/opre.16.5.955
- TRANG TRAN, T. N., ATAS, M., Felfernig, A., & Stettinger, M. (2017). An overview of recommender systems in the healthy food domain. *Journal of Intelligent Information Systems*. doi: 10.1007/s10844-017-0469-0
- UNGAR, L., & FOSTER, D. P. (1998). A formal statistical approach to collaborative filtering. *CONALD '98*.

- YANG, J., WANG, W., WANG, H., & YU, P. (2002). *δ -clusters: capturing subspace correlation in a large data set*. In: Data Engineering, 2002. Proceedings. 18th International Conference on.
- ZHANG, M., & HURLEY, N. (2008). *Avoiding monotony: improving the diversity of recommendation lists*. In: Proceedings of the 2008 ACM conference on Recommender systems.
- ZHENG, K., *et al.* (2017). A survey of query result diversification. *Knowledge and Information Systems*, 51(1), 1-36. doi: 10.1007/s10115-016-0990-4
- ZHOU, T., *et al.* (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10), 4511-4515. doi: 10.1073/pnas.1000488107
- ZHOU, T., *et al.* (2009). Accurate and diverse recommendations via eliminating redundant correlations. *New Journal of Physics*, 11(12), 123008. doi: 10.1088/1367-2630/11/12/123008
- ZIEGLER, C.-N., MCNEE, S. M., KONSTAN, J. A., & LAUSEN, G. (2005). *Improving recommendation lists through topic diversification*. In: Proceedings of the 14th international conference on World Wide Web.