Paper 1886-2014

Recommending News Articles using Cosine Similarity Function

Rajendra LVN[1], Qing Wang[2] and John Dilip Raj[1]
[1]GE Capital Retail Finance, [2]Warwick Business School

ABSTRACT

Predicting news articles customers are likely to view/read next provides a distinct advantage to news sites and Collaborative filtering is a widely used technique for the same. This paper details an approach within Collaborative Filtering that uses the 'Cosine Similarity Function' to achieve this purpose. The paper further details two different approaches, customized targeting and article level targeting that can be used in marketing campaigns.

INTRODUCTION

All through history, people have relied on some kind of observation/advice/feedback in making decisions of any kind. With information in the web increasing manifold and various options to choose from, customers at times find it difficult to search and read articles that are of most interest to them. News sites have stepped in to fill this gap by analyzing customer behavior and recommending articles that customers have high likelihood to read. Drawing from fields like cognitive science (Rich 1979) and information retrieval (Mooers 1950) and with great advances in technology, recommending articles consumers are most likely to read has fast grown to be an integral part of many marketing and CRM processes (Huang, Zeng and Chen 2007).

COLLABORATIVE FILTERING

There are many approaches that can be used to recommend articles to customers. Huang, Chung and Chen (2004) compare four such approaches – knowledge engineering, collaborative filtering, content-based and hybrid. Collaborative filtering is widely believed to be one of the most successful recommendation approaches. This method relies on the past viewership behavior alone to recommend articles to customers. Su and Khoshgoftaar (2009) provide a survey of collaborative filtering techniques. These take into consideration only past viewed history and not the attributes of the customer or article.

User-based collaborative filtering has been used extensively for building recommendation systems. However, with larger datasets these systems become more computationally complex to build. Item-based techniques, where user-item matrix multiplication is used to find similarity between different items, address these scalability issues. Karypis (2001) estimates item-based algorithms to be 28 times faster and more importantly the recommendations to be 27% better than traditional user-based algorithms.

COSINE SIMILARITY FUNCTION

One of the popular approaches in collaborative filtering is finding similarities using the cosine similarity function, which is the measure of similarity between two vectors derived from the cosine of the angle between them. To find the relationship between two news articles, each article is treated as a vector in the space of consumers. The cosine between these vectors gives a measure of similarity. Such functions have largely been used in the web space to identify similarity of text documents and web pages. It has been one of the most preferred techniques in information retrieval, clustering and even applied to pattern recognition and medical diagnosis (Ye 2011).

To put this in practice, we take the algorithms detailed by Deshpande and Karypis (2004) and convert them as SAS® codes to achieve the same.

Step 1: Input the transaction level dataset

A simple example to illustrate the technique is used here. Table 1 shows 5 customers and their reading behavior on 4 articles in a given time period.

Only a binary value as to whether a customer has read the article in the recent past or not is considered. Whether customers read an article multiple times is not taken into account, consistent with research (Deshpande and Karypis 2004) that shows higher recommendation performance using this method.

Table 1: Customer – News Article Matrix

Customer ID	Article 1	Article 2	Article 3	Article 4
1	0	1	0	1
2	0	1	1	1
3	1	0	1	0
4	1	0	0	1
5	1	0	0	0

Step 2: Calculate News Article Similarity using Cosine Angle

To find the similarities between the above articles, we use the approach proposed by Deshpande and Karypis (2004) where cosine similarities of every combination of 2 articles at a time are done (n-1)!. This can be done in SAS® using PROC IML in the following way.

```
PROC IML;
use data;
read all var{m_1} into p1;                  /* input Vector 1 ex: m_1  */
read all var{m_3} into p2;                  /* input Vector 2 ex: m_3  */
wgts=p1`//p2`;                              /* to create a 2X5 Matrix  */
        norm = sqrt(wgts[ ,##]);            /* euclidean norm          */
        a = wgts/norm;
        d = a*a`;
/*print wgts norm a d;*/
create out from d;
append from d;
QUIT;
```

The same can also be done without PROC IML and completely using SQL, but for the purpose of this paper, the former approach is detailed. Table 2 gives the article similarity matrix based on the data we have from Table 1.

Table 2: News Article Similarity Matrix

	Article 1	**Article 2**	**Article 3**	**Article 4**
Article 1	0	0	0.41	0.33
Article 2	0	0	0.50	0.82
Article 3	0.41	0.50	0	0.41
Article 4	0.33	0.82	0.41	0

There are a couple of points to note here. One, the diagonal is inputted with value ‘0’ (reason detailed later in the paper). Two, is to understand articles that are most similar and dissimilar. News articles 2 and 4 are most similar with a value of 0.82 because 4 out of 5 customers displayed the same behavior towards these. Two customers viewed these articles while two others did not view them. Only one customer viewed one of the articles and not the other. Therefore, 4 of 5 customers displayed same behavior, either viewing them or not viewing them. The most dissimilar articles are 1 and 2. This is because all 5 customers viewed either of the two articles but not both.

Step 3: Targeting customers using recommendation strength of articles

Table 3 is the news article recommendation strength rolled up at customer level.

Table 3: Recommendation Strength of News Articles at Customer Level

Customer	Article 1	Article 2	Article 3	Article 4
1	0.33	0.82	0.91	0.82
2	0.74	1.32	0.91	1.22
3	0.41	0.50	0.41	0.74
4	0.33	0.82	0.82	0.33
5	0	0	0.41	0.33

Let's take the example of one customer and see how the recommendation strengths for a news article are arrived at. From the above table, we see Article 3 is the one that Customer 1 is most likely to read. This is in spite of the customer not viewing that article in the time period on which the cosine similarity was built. Let us see how Article 3 gets a recommendation strength of 0.91 for the customer. The first step is to identify articles viewed by the customer in the observation time period. These would be Articles 2 and 4 as shown in Table 1. Now, the similarities between Article 3 and articles viewed by the customer earlier (Articles 2 and 4) are known from Table 2. Their sum (0.5 + 0.41) gives the Recommendation Strength for Article 3 for this customer. The other article (1) not viewed by the customer however gets the lowest recommendation strength, using the same calculation (0 + 0.33).

Different Approaches in Targeting

Now that we know which articles customers are likely to view, the next step is to develop ways of targeting customers. We present two ways of targeting.

Step 4a: Customized targeting (Method 1)

Now that we have the recommendation strengths of each article at customer level, we can market articles to customers based on the recommendation strengths. Table 4 details the same. As there is article-wise ranking, the optimum number of recommendations can be made, be it 2, 3 or any number.

Table 4: Customer Targeting by Recommendation Strength

Customer ID	First Article Preference	Second Article Preference	Third Article Preference	Fourth Article Preference	Strength of First Article Preference	Strength of Second Article Preference	Strength of Third Article Preference	Strength of Fourth Article Preference
1	3	2	4	1	0.91	0.82	0.82	0.33
2	2	4	3	1	1.32	1.22	0.91	0.74
3	4	2	1	3	0.74	0.5	0.41	0.41
4	3	2	1	4	0.82	0.82	0.33	0.33
5	3	4	.	.	0.41	0.33	.	.

We see for Customer 5, there are only two news articles that are recommended. This is because the customer viewed only Article 1, which is similar to only Articles 3 and 4. Therefore these two articles are recommended while Article 2 is not recommended because Articles 1 & 2 have not been viewed by any customer as shown in Table 1. Another point to note is that if only one article has been viewed by the customer, as was the case for Customer 5, the same article is not recommended because the cosine similarity between the same articles is 1 and if this value is inputted, it may unduly dominate & influence other news article recommendations.

Step 4b: Article level targeting (Method 2)

In an ideal scenario, customized targeting as detailed in the previous step can be done. At times though, the business objective may be to promote a given article. Customized targeting approach though for the same in a real environment may be difficult because of the numerous creative that have to be designed to customize each article. For this purpose, it is better to select certain news articles (and not all) where viewership need to be increased and target customers who are most likely to view in the same. Table 4 provides the order in which customers would be targeted based on the articles chosen, using Cosine Similarity in marketing campaigns.

Table 4b: Customer Targeting by News Article

Article	First Customer to Target	Second Customer to Target	Third Customer to Target	Fourth Customer to Target	Fifth Customer to Target
1	2	3	1	4	5
2	2	1	4	3	5
3	1	2	4	3	5
4	2	1	3	4	5

For example, if the business mandate is to increase viewership in Article 4, the order of targeting can be in decreasing order of the recommendation strength within that article. If only the top 60% of the customers are to be chosen this way, Customer 2 will be chosen first, followed by Customer 1 and then Customer 3.

CONCLUSION and FUTURE DIRECTIONS FOR RESEACH

Cosine Similarity is a powerful way to predict customer preferences towards news articles and is an invaluable tool. The approach suggested in this paper uses only binary level data for each news article and therefore is also easy to build and implement.

REFERENCES

Deshpande, M., & Karypis, G. (2004), 'Item-Based Top-N Recommendation Algorithms', *ACM Transactions on Information Systems*, 22(1), 143-177.

Huang, Z., Chung, W. & Chen, H. (2004), 'A Graph Model for E-Commerce Recommender Systems', *Journal of the American Society for Information Science and Technology*, 55(3), 259-274.

Huang, Z., Zeng, D. D. & Chen H. (2007), 'Analyzing Consumer-Product Graphs: Empirical Findings and Applications in Recommender Systems', *Management Science*, 53(7), 1146-1164.

Karypis, G. (2001), 'Evaluation of Item-Based Top-N Recommendation Algorithms', in *Proceedings of the ACM Conference on Information and Knowledge Management*. ACM, New York, 247-254.

Rich, E. (1979), 'User Modeling via Stereotypes', *Cognitive Science*, 3(4), 329–354.

Mooers, C. N. (1950), 'The Theory of Digital Handling of Non-Numerical Information and Its Implications to Machine Economics. Technical Bulletin No. 48. Cambridge, MA: Zator Co., 1950 (Paper presented at the Association for Computing Machinery, Rutgers Univ., New Brunswick, NJ, 1950 March 29).

Su, X. & Khoshgoftaar, T. M. (2009), 'A Survey of Collaborative Filtering Techniques', *Advances in Artificial Intelligence*, Article ID 421425, 19 pages.

SAS Institute Inc. "Introduction to SAS/IML Software", SAS/IML(R) 9.3 User's Guide. http://support.sas.com/documentation/cdl/en/imlug/64248/HTML/default/viewer.htm#imlstart_toc.htm

Ye, J. (2011), 'Cosine Similarity Measures for Intuitionistic Fuzzy Sets and their Applications', *Mathematical and Computer Modelling*, 53(1-2), 91-97.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Rajendra LVN
GE
iLabs Building, Block 1A, Hitech City,
Madhapur, Hyderabad – 500081
India
Email: rajendra.ledallavenkatanaga@ge.com