

Empirical Evaluation of Content-Based Filtering for Personalization

*Joshua B. Hurwitz**

User-Centered Solutions Lab – Motorola Labs, Schaumburg, Illinois, USA

joshua.hurwitz@motorola.com

Abstract

With the increasing prevalence of services that provide content on mobile devices, it will become necessary to simplify the process of searching for content using these devices. One potential solution is to use Content-based filtering to personalize searches. In this approach, users store their interests in personal profiles on their devices, and an intelligent system finds content which has metadata that matches these interests. However, one assumption of this method is that the terms used in the profile accurately reflect what makes users interested in the content. The current study evaluated this assumption by analyzing the data of survey respondents who rated their interests in topics and articles in 6 domains of news. The results showed differences across these domains in the consistency and accuracy of predictions of content interests. Consistency was the highest for the Science-Technology domain and the lowest for Business, Entertainment and Health, and accuracy was lowest for Entertainment. These results are discussed in terms of their implications for metadata standards such as TV Anytime.

Key words: mobile services, content, metadata, personal profiles, user interfaces

1. General

Telecommunications services are expanding to include distribution of content to users on handheld mobile devices. For example, Motorola recently introduced its SCREEN3 service, which provides access to news, sports, entertainment and other content on a mobile device [Rubin (2005)]. Also, The New York Times has teamed up with Vindigo Studios to provide full-length news articles to Verizon customers on their mobile devices [m-Travel.com (2003)] and the Google search engine has been made available to AT&T wireless customers [m-Travel.com (2001)].

While these services expand the opportunities for mobile device users, the human interfaces of mobile devices—the small screens and limited keypads—make it difficult to perform the types of large-scale searches that can be performed on personal computers [Buchanan, Farrant, Jones, & Thimbleby (2001)]. Consequently, it is more difficult for users to find their preferred content when they use a mobile device than when they use a personal computer. This problem is further exacerbated as the volume of content on the Internet grows.

One solution has been to use automated personalized search mechanisms that employ Content-based filtering [e.g., Loeb (1992)]. With this approach, users store information regarding their content preferences in personal profiles on their handheld devices, and the

* The author would like to thank Thea Turner, Miles Jackson, Paola Hobson and Genevieve Conaty for their contributions to this research.

personalization system selects content by matching these preferences with metadata representing the content. Thus, the majority of the work performed by the user is to set up the initial profile. While the user may manually update the profile with changes or additions to personal preferences, some systems automate this process, learning the user's preferences from their content choices [e.g., Carreira, Crato, Gonçalves, & Jorge (2004)].

1.1 Problems with Content-based Filtering

One assumption of this approach is that user content preferences can be predicted from their preferences for content categories. For example, users who state a preference for reading articles about dance should prefer articles that discuss dancers, dance troupes and dance performances. However, this assumption has not been empirically tested across a variety of content domains. Furthermore, it relies on the presumption that the system for categorizing content reflects the user's mental representation of that content. Thus, the user's definition of a category label would ideally have to be the same as the definition used by the individuals who designed the metadata system for categorizing the content.

A further complication is that efforts to improve the accuracy of the system may contradict other goals of adding personalization to mobile devices. For example, attempting to increase accuracy by adding more category terms for use in the personal profile and in the metadata will increase the work required by users to set up their profiles. It could also, paradoxically, increase the likelihood of error, since there are more opportunities for erroneous mismatches between the metadata and the terms stored in the personal profile.

1.2 Factor Analytic Approach

Given these issues, the goal in developing a corpus of metadata terms should be to identify an optimal categorization scheme that best represents the full space of the content, as it is understood by the users, with the fewest terms possible. In such a scheme, the terms should be maximally dissimilar from each other, so as to minimize redundancy in the categories. This means that user interest in each category in a domain should ideally be unrelated to their interest in other categories in that domain. Thus, each category would provide a unique contribution to predicting user interest in content in that domain.

One statistical method for achieving this category structure is factor analysis. Consider the case of news content, which is used in the current study. Given a set of user interest ratings for various topics in a news domain (e.g., Science-Technology), factor analysis identifies factors consisting of topics with similar ratings. For each topic that "loads" on a factor, the ratings for that topic correlate more highly with the ratings for other topics that also load on that factor than they do with ratings for topics in other factors. Thus, factor analysis can be used to identify groups of news topics within a domain, with each group consisting of topics that users appear to treat as nearly equally interesting or uninteresting to them. The resulting factor structure could then reduce the space of topics down to a more manageable set, since it identifies topics that are redundant with each other in terms of user interest.

Given this factor structure, factor analysis also computes a summary score, called the factor score, for each factor and for every respondent. In some versions of factor analysis, the vector of scores for each factor is orthogonal to the vector for the other factors, so that the scores for each factor are statistically uncorrelated with those of the other factors. Each vector thus represents the factor's unique contribution to the respondents' interest ratings in that domain.

1.3 Predicting Interest in Content

If, for a given domain, there is a positive relationship between 1) user interest in a set of similar news topics and 2) user interest in related articles in that domain, then there should be a positive correlation between 1) the factor scores representing the ratings of those topics and 2) the ratings of the news articles. Furthermore, this relationship should be observed across different samples of users. Such reliability is important if user interest in news topics is to be used by a personalization system as an accurate indicator of interest in the content.

Thus, given the need for efficient and reliable interest descriptors that help to accurately predict user interests in content, the current study analyzed data collected from two surveys on user preferences for news topics and content. Each survey asked respondents about their interests in a variety of news topics and articles. The goal was to assess whether an efficient set of descriptors of user interest could help to reliably select content of interest to users.

2. Method

2.1 Subjects

Data were collected from two sets of respondents. In Study 1, there were 681 Motorola employees, 468 males and 213 females, ranging in age from 18 to 65 years, with 59% over 35 years and 35% between 25 and 34 years. In Study 2, there were 273 college students, 227 males and 46 females, ranging in age from 16 to 34 years, with 71% between 16 and 21 years and 15% between 22 and 24 years. The Motorola employees reported living in 28 countries, with 46.1% from North America, followed by 32.8% from European countries and 13.4% from countries in Asia. The students in the second study attended 13 universities spread across 8 countries: Greece, the UK, Spain, Portugal, Italy, the Czech Republic, the US, and India. The vast majority of these students, about 85%, reported majoring in a technical subject in school, either in Engineering, Computer Science, the Physical Sciences or Mathematics.

2.2 Procedure

The survey was administered over the Internet through SurveyMonkey (www.surveymonkey.com). It consisted of questions in a range of areas, including 1) respondents' demographics (gender, age, etc.), 2) their activities relating to consumption of news, movie entertainment and television programs, 3) their preferences for content alerts (i.e., alerts to help screen content not suitable for children) and preferences for types of television entertainment based on mood, 4) their interests in news topics, and 5) their preferences for news content. The results presented here only address the last two areas.

In the section on news topics, the respondents were presented with topics in 6 domains: Business-Finance, Entertainment, Government-Politics, Health, Sports and Science-Technology. There were 45 topics in each domain, which were selected to cover the scope of the domain as much as possible. For each topic, the respondents were asked to "rate how interested [they] would be in reading a news item on that topic". They gave their responses on a scale from 1 to 5, where 1 meant "Not at all Interested", 3 meant "Somewhat Interested" and 5 meant "Extremely Interested".

In the section on news content, there were 15 articles in each domain. The articles in each domain were derived from on-line sources, and together covered a wide range of topics within

that domain. Each article was edited so that its length varied from 200 to 270 words, to allow for reading times of no more than 1.5 minutes at normal reading rates on computer screens [Mills and Weldon (1988)].

For each article, respondents were given the following instructions: “Please read the short news item below, then rate how interested you are in that item. If you have already read or heard something about the subject, please answer as if you had not.” After they read the article, they gave their response on the same scale used for rating the news topics.

3. Results

3.1 Topic Ratings

The topic ratings in each domain of news were analyzed using Principal Components Analysis with Varimax rotation. The number of factors in each analysis was determined by the size of the eigenvalue associated with each factor, with factors being accepted only if their eigenvalues were greater than or equal to 1. Also, the analysis produced Anderson-Rubin factor scores for each factor, which ensured that the scores varied between 0 and 1 and that the vectors of scores for the estimated factors were orthogonal to each other. Finally, the threshold for deciding that a topic loaded on a factor was whether the correlation coefficient between the topic ratings and the factor scores exceeded 0.5.

Table 1 shows the means and standard deviations of the topic ratings, along with the number of factors and percentage of variance accounted for by each factor analysis in each of the two studies. In both studies, the average ratings were the highest in Science-Technology and the lowest in Sports. Also, in both studies, the analyses for Entertainment and Sports produced the most factors (7 each), whereas in Study 1, the analyses of Business-Finance and Government-Politics produced the fewest factors (4 each). In Study 2, the analysis of the Business-Finance domain also produced the fewest factors (5).

Table 1. Basic statistics for the news topic ratings in both studies.

Category	Averages (SD)		Number of Factors		Percentage of Variance	
	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
Business-Finance	2.65(0.86)	2.27(1.30)	4	5	67%	69%
Entertainment	2.24(0.62)	2.49(1.42)	7	9	61%	67%
Government-Politics	2.57(0.82)	2.33(1.31)	4	6	67%	67%
Health	2.38(0.81)	2.37(1.40)	6	6	65%	70%
Sports	1.86(0.68)	2.06(1.40)	7	7	65%	67%
Science-Technology	2.69(0.81)	2.95(1.60)	5	6	66%	67%

Table 2 shows the highest and second-highest overlap between the Study-1 and Study-2 factors. If U_{ij} is the number of topics in a domain that load on either factor i in Study 1 or factor j in Study 2 (i.e., the union of the two factors), and I_{ij} is the number of topics that load on both i and j (i.e., their intersection), then the overlap, T_{ij} , is I_{ij}/U_{ij} . This metric is an indicator of the consistency of the factor solutions across the two studies. If a factor analysis in a domain in the second study perfectly replicated the results from the first study, then the highest overlap for all of the factors in that domain would be 1, and the second-highest overlap would be 0.

Using this metric, Table 2 shows that the domains of Entertainment, Health and Science-Technology had the highest consistency across the two studies. For most of the factors in these domains, the highest overlap was at least 0.5, and the second-highest was close to or equal to 0. The exceptions were the Study-1 factors 4 and 7 in Entertainment, the Study-1 factor 3 in Health, and the Study-1 factor 3 in Science-Technology.

Business-Finance, Government-Politics and Sports had the lowest overlap values. In Business-Finance and Government-Politics, half of the Study-1 factors had maximum overlap values that were below 0.5. In the Sports domain, these values for 5 of the 7 Study-1 factors were below 0.5, and 2 of these had no overlap with any of the Study-2 factors.

Table 2. Overlap between Study-1 and Study-2 factors.

Study-1 Factor	Bus		Ent		Hea		Gov		Spo		Sci	
	Highest Overlap	2nd-Highest Overlap	Highest Overlap	2nd-Highest Overlap	Highest Overlap	2nd-Highest Overlap	Highest Overlap	2nd-Highest Overlap	Highest Overlap	2nd-Highest Overlap	Highest Overlap	2nd-Highest Overlap
1	0.53 (1)	0.05 (3)	0.67 (1)	**	0.75 (1)	**	0.79 (1)	**	0.55 (1)	**	0.93 (1)	**
2	0.36 (2)	0.33 (5)	0.71 (4)	0.08 (3)	0.83 (3)	**	0.13 (3)	0.07 (1)	0.29 (5)	0.15 (1)	0.86 (3)	0.07 (2)
3	0.25 (3)	0.09 (1)	1.00 (2)	**	0.43 (4)	0.08 (1)	0.14 (6)	**	0.25 (7)	0.06 (1)	0.44 (5)	0.38 (2)
4	0.60 (4)	0.09 (2)	0.43 (5)	0.17 (7)	0.63 (2)	**	1.00 (5)	**	0.43 (3)	0.40 (6)	0.78 (4)	0.06 (2)
5			0.83 (3)	**	0.50 (5)	**			*	**	1.00 (6)	**
6			1.00 (6)	**	1.00 (6)	**			*	**		
7			0.20 (5)	**					0.67 (4)	**		

Note that the Study-2 factors are indicated inside parentheses next to each overlap value.

* Indicates that there was no overlap in topics between the Study-1 factor and any Study-2 factor.

** Indicates that there was no overlap in topics between the Study-1 factor and any Study-2 factor besides the one with the highest overlap.

3.2 Predicting Article Ratings

Table 3 shows the means and standard deviations for the article ratings, as well as the number of articles for which the ratings were significantly predicted by at least 1 factor. A prediction of an article's ratings by a factor was considered significant if the *t*-test of the correlation between the factor scores and the ratings was significant at $p < .001$, and if the percentage of variance of the article ratings accounted for by the factor scores exceeded 10%. Note that for Study 2, 81 respondents did not complete the article ratings, so only the results from the remaining 192 respondents are presented here.

As with the topic ratings, the averages for the article ratings in both studies were the highest for Science-Technology articles, and were the lowest for Sports articles. Also, despite the relatively large number of factors in the Entertainment domain (see Table 1), there were relatively few articles for which there were significant predictions in that domain.

In addition to looking at the number of significant predictions for each domain, the analyses addressed the question of whether consistency in the factors across the two studies produced consistency in predicting the article ratings as well. In other words, if the topics that loaded on a factor in Study 1 were mostly the same as those that loaded on a factor in Study 2, then there would presumably also be a large overlap in the articles that these two factors predict.

One way to evaluate this is to examine the Root Mean Squared Deviation (RMSD) between the overlap scores, T_{ij} , for the topic factors, and the analogous scores, A_{ij} , for predicting the articles. If V_{ij} is the number of articles in a domain that are significantly predicted by factor i in Study 1 or factor j in Study 2 (i.e., the union of the factors), and J_{ij} is the number of articles

that are significantly predicted by both i and j (i.e., their intersection), then the overlap, A_{ij} , is J_{ij}/V_{ij} . If N_1 is the number of factors in Study 1 and N_2 is the number in Study 2, then

$$\text{RMSD} = \left(\frac{\sum_i \sum_j (T_{ij} - A_{ij})^2}{N_1 N_2} \right)^{0.5}$$

The right column in Table 3 shows the RMSDs for the 4 domains. The results show that Science-Technology had the lowest RMSD whereas Business-Finance, Health, and Entertainment had the largest RMSDs. In Science-Technology, the same 14 topics loaded on the first factor—Computers—in each study. These topics included, for example, “Computer Hardware”, “Robotics”, and “Nanotechnology”. Furthermore, this factor significantly predicted ratings for the same 8 articles in both studies. These included articles on the Internet search engine Yahoo!, on electronic paper, and on the Chandra Space Telescope. Similar consistencies were observed with the other factors, although they were not as consistent as the first factor.

In Health, the 4th factor in Study 1, labeled “Baby/Child Health” and the 2nd factor in Study 2, called “Parenting/Reproductive Health”, had a topic overlap of 0.63, since they shared the same 5 topics (e.g., “Approaches to Parenting” and “Children’s Health”), but two additional topics loaded on the Study-2 factor. However, despite this overlap, the factor scores for the Study-1 factor produced significant correlations with the ratings for 1 article on children’s diets, whereas the scores for the Study-2 factor produced significant correlations with 4 articles on elderly health, HIV and breast cancer.

A similar pattern was observed in Entertainment. The 5th factor in Study 1 and the 3rd in Study 2 (both labeled “Theatre/Art”) shared the same 5 topics, except that the Study-2 factor included an additional topic (“Improvisational Theatre”). Despite this overlap, the Study-1 factor did not significantly predict ratings for any articles, whereas the Study-2 factor predicted ratings for articles on the author Sir Arthur Conan Doyle, the dance troupe “Riverdance”, and the Shakespeare play “King Lear”.

Table 3. Basic Statistics for the article ratings, and for predicting the article ratings.

Category	Averages (SD)		Number Sig. Predictions		RMSD
	Study 1	Study 2	Study 1	Study 2	
Business-Finance	2.67(1.17)	2.66(1.09)	13	10	0.44
Entertainment	2.28(1.15)	2.64(1.11)	8	11	0.25
Government-Politics	2.88(1.18)	2.80(1.06)	15	13	0.23
Health	2.49(1.17)	2.60(1.10)	15	12	0.34
Sports	1.79(1.08)	2.20(1.18)	15	14	0.20
Science-Technology	3.28(1.19)	3.54(1.15)	13	14	0.17

4. Discussion

These results demonstrate that interest in content categories may not necessarily be a reliable indicator of interest in the content. Some news content domains, such as Science-Technology, showed very high consistency in the factor structure and the prediction of article ratings. Predictions in the Business and Health domains were less reliable, so that the same

articles were not necessarily predicted by the same factors in the two studies. Predictions in the Entertainment domain were lower than in the other domains, with as many as 7 out of the 15 articles not being predicted by any of the factors.

These results suggest that Content-based filtering for mobile content services may not produce the best results when the system selects content to download to a user's mobile device. The terms stored in the user's personal profile may not adequately describe their interests in content, so they may be presented with content that is not interesting to them and may miss out on content that they do find appealing. If these errors occur too often, then user interest in such systems could decline as they lose trust in the system's ability to choose the appropriate content for them [Lee, Kim, & Chung (2002)].

The differences in the magnitude and reliability of predictions that were observed across different domains of news may have been due to the nature of the terms in these domains. Terms in the Science-Technology domain are likely to be more concrete than those in some of the other domains, such as Entertainment. So users may be more uniform in their definitions of topics such as "Computer Hardware" and "Desktop Publishing" than they are in defining terms such as "Dramatic Television Series" and "Music" that describe creative content.

Thus, when confronted with an article in the Science-Technology domain, as opposed to an article in Entertainment, there would be a tighter link between the terms used in the article and the terms in the user profile that describe the user's interests. For example, most users would probably agree that an article that describes a search engine is relevant to their interests in the Internet. However, while an article on the group "Riverdance" may be interesting to users who want to read about the topic of dance, it may also be interesting to other users because they consider Riverdance to be a popular and entertaining group.

Another explanation for these differences may have been cultural and age-related variations in interpreting the topic descriptors. Study 1 sampled from Motorola employees who were generally older than the respondents in Study 2. This may, for example, have contributed to differences between the two studies in results for the Business-Finance domain. Compared to the students, many of the Motorola employees were more likely to have had real-life experience in this domain, including studying the marketplace and dealing with personal finances. This greater experience means that Motorola employees were probably more likely to agree on the definitions of topics such as "Business Finance" and "Resource Management".

The reason for the consistency in factors and article predictions in the Science-Technology domain may have been that the majority of respondents in both studies had an interest in science- and technology-related subjects. The majority of respondents in Study 2 were students who reported that they majored in a subject in these areas. Furthermore, while the Motorola employees in Study 1 were not asked about their career specialties, it is likely that most of them had some background in engineering or related disciplines. This likely common experience across the two studies may have contributed to the consistency in both the topic ratings and the predictions of content ratings across the studies.

4.1 Implications for Metadata Standards

In the past, Metadata Standards such as TV Anytime have proposed categorization schemes for defining a corpus of content descriptors that may not necessarily reflect how users categorize content. These schemes could benefit from analyses of empirical user data such as

the ones performed in the current study. The factor analytic approach taken here could help to improve the efficiency of metadata by identifying groups of topic terms that appear to be similar in terms of user interests. An efficient metadata system could then use the most representative terms in these groups (i.e., those that load most highly on their respective factors) as descriptors for the personal profile and metadata.

However, even using this approach, there are other weaknesses in relying on Content-based filtering to select content for users. Culture, age, experience and other individual-differences variables could produce variations in how users interpret topic terms in those domains. The lack of concreteness in topic definitions for certain domains (e.g., creative content) could also reduce the predictability of user interest in content, with different users employing different definitions for the topic terms.

The solution may be to rely on Collaborative filtering or some combination of Collaborative and Content-based filtering to identify relevant content for users [Paulson & Tzanavari (2003)]. With this approach, content would be selected for a user according to whether the metadata associated with that content matches the user's interests and whether users with a similar configuration of interests have preferred that content. The personalization system might then rely more heavily on one filtering approach or the other depending on the domain of the content. Thus, combining both more efficient metadata with a hybrid filtering system would reduce user work in setting up their personal profile on a mobile device while helping to improve accuracy in predicting users' interests in the content.

References

- Buchanan, G., Farrant, S., Jones, M., & Thimbleby, H. (2001). *Improving Mobile Internet Usability*. Paper presented at the WWW10, Hong Kong.
- Carreira, R., Crato, J. M., Gonçalves, D., & Jorge, J. A. (2004). *Evaluating adaptive user profiles for news classification*. Paper presented at the 9th International Conference on Intelligent User Interfaces, Funchal, Madeira, Portugal.
- Lee, W. J., Kim, T. U., & Chung, J.-Y. (2002). *User acceptance of the mobile Internet*, Proceedings of the First International Conference on Mobile Business. Athens, Greece: Mobiforum.
- Loeb, S. (1992). Architecting personalized delivery of multimedia information. *Communications of the ACM*, 35(12), 39-47.
- m-Travel.com. (2001). *Google provides searches for AT&T wireless*. Available: http://www.m-travel.com/news/2001/10/google_provides.html, February 15, 2006.
- m-Travel.com. (2003). *New York Times offered for mobile devices*. Available: http://www.m-travel.com/news/2003/10/new_york_times_.html, February 15, 2006.
- Paulson, P., & Tzanavari, A. (2003). Combining collaborative and content-based filtering using conceptual graphs. *Lecture Notes in Computer Science*, 2873, 168-185.
- Rubin, J. (2005). *Motorola Screen3*. Available: http://www.coolhunting.com/archives/2005/03/motorola_screen.php, February 15, 2006.