User Evaluation of Multidimensional Relevance Assessment

Célia da Costa Pereira
Università degli Studi di Milano
Dipartimento di Tecnologie dell'Informazione
Via Bramante 65, I-26013
Crema (CR), Italy
pereira@dti.unimi.it

Mauro Dragoni
Università degli Studi di Milano
Dipartimento di Tecnologie dell'Informazione
Via Bramante 65, I-26013
Crema (CR), Italy
dragoni@dti.unimi.it

Gabriella Pasi
Università degli Studi di Milano Bicocca
Dipartimento di Informatica Sistemistica e Comunicazione
Viale Sarca, 336, I-20126
Milano (MI), Italy
pasi@disco.unimib.it

ABSTRACT

In this paper a user evaluation is proposed to assess the effectiveness of systems based on multidimensional relevance assessment. First of all, we introduce our approach to multidimensional modeling and aggregation, and the criteria used for the experiments. Then, we describe how the user evaluation has been performed, and finally, we discuss the results obtained.

1. INTRODUCTION

In the first traditional approaches to Information Retrieval (IR), relevance was modeled as "topicality", and its numeric assessment was based on the matching function related to the adopted IR model (*boolean model*, *vector space model*, *probabilistic model* or *fuzzy model*). However, relevance is, in its very nature, the result of several components or dimensions. Cooper [2] can be considered as one of the first researchers who had intuitions on the multidimensional nature of the concept of relevance. He defined relevance as *topical relevance with utility*. Mizzaro, who has written an interesting article on the history of relevance [8], proposed a relevance model in which relevance is represented as a four-dimensional relationship between an information resource (surrogate, document, and information) and a representation of the user's problem (query, request, real information need and perceived information need). A further judgment is made according to the: topic, task, or context, at a particular point in time. The dimensions pointed out by Mizzaro are in line with the five manifestations of relevance suggested by Saracevic [10]: *system or algorithmic relevance*, *topical or subject relevance*, *cognitive relevance or pertinence*, *situational relevance or utility* and *motivational or effective relevance*. However, the concept of *dimension* used in this paper which is similar to that used by Xu and Chen in [12] is somehow different from that used by Mizzaro and Saracevic. They defined several kinds of relevance and call them *dimensions of relevance* while we define relevance as a *concept of concepts*, i.e., as a point in a n-dimensional space composed by n criteria. The document score is then the result of a particular combination of those n space components as explained in [3, 4].

One of the problems raised by considering relevance as a multidimensional property of documents is how to aggregate the related relevance scores. In [3, 4] an approach for prioritized aggregation of multidimensional relevance has been proposed. The proposed aggregation scheme is user dependent: a user can be differently interested in each dimension. The computation of the overall relevance score to be associated with each retrieved document is then based on the aggregation of the scores representing the satisfaction of the considered dimensions. A problem raised by this new approach is how to evaluate its effectiveness. In fact, there is no test collection suited to evaluate such a model. In this paper, we first recall the models for aggregating multiple dimensions evaluations for relevance assessment presented in [3] and [4]. We focus on observing how document rankings are modified after applying the two operators on the different typologies of users (different dimensions orderings).

The paper is organized as follows. Section 2 recalls the aggregation models used in the paper. Section 3 presents the performed user evaluation and, finally, Section 4 concludes the paper.

2. PRIORITIZED MULTICRITERIA AGGREGATION

In this section, after a brief background on the representation of a multicriteria decision making problem, two prioritized approaches for aggregating distinct relevance assessments are shortly presented.

2.1 Problem Representation

The presented multicriteria decision making approaches have the following components:

- the set C of the n considered criteria: $C = \{C_1, \ldots, C_n\}$, with C_i being the function evaluating the ith criterion;
- the collection of documents D;
- an aggregation function F to calculate for each document $d \in D$ a score $F(C(d))$[1] $= RSV(d)$ on the basis of the evaluation scores of the considered criteria.

[1]Actually, it corresponds to $F(C_1(d), \ldots, C_n(d))$.

Appears in the Proceedings of the 1st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padova, Italy.
http://ims.dei.unipd.it/websites/iir10/index.html

$C_j(d)$ represents the satisfaction scores of document d with respect to criterion j. The weight associated with each criterion $C_i \in C$, with $i \neq 1$, is document and user-dependent. It depends on the preference order of C_i for the user, and also on both the weight associated to criterion C_{i-1}, and the satisfaction degree of the document with respect to C_{i-1}[2]. Formally, if we consider document d, each criterion C_i has an importance $\lambda_i \in [0, 1]$.

Notice that different users can have a different preference order over the criteria and, therefore, it is possible to obtain different importance weights for the same document for different users.

We suppose that $C_i \succ C_j$ if $i < j$. This is just a representational convention which means that the most preferred criteria have lower indexes.

We suppose that:

- for each document d, the weight of the most important criterion C_1 is set to 1, i.e., by definition we have: $\forall\, d\ \lambda_1 = 1$;
- the weights of the other criteria C_i, $i \in [2, n]$, are calculated as follows:

$$\lambda_i = \lambda_{i-1} \cdot C_{i-1}(d), \tag{1}$$

where $C_{i-1}(d)$ is the degree of satisfaction of criterion C_{i-1} by document d, and λ_{i-1} is the importance weight of criterion C_{i-1}.

2.2 The Prioritized Scoring model

This operator allows us to calculate the overall score value from several criteria, where the weight of each criterion depends both on the weights and on the satisfaction degrees of the most important criteria — the higher the satisfaction degree of a more important criterion, the more the satisfaction degree of a less important criterion influences the overall score.

Operator F_s is defined as follow: $F_s : [0, 1]^n \rightarrow [0, n]$ and it is such that, for any document d,

$$F_s(C_1(d), \ldots, C_n(d)) = \sum_{i=1}^{n} \lambda_i \cdot C_i(d). \tag{2}$$

The RSV_s of the alternative d is then given by:

$$RSV_s(d) = F_s(C_1(d), \ldots, C_n(d)). \tag{3}$$

Formalizations and properties of this operator are presented in [3].

2.3 The Prioritized "min" Operator

In this section a prioritized "min" (or "and") operator is recalled [4]. This operator allows to compute the overall satisfaction degree for a user whose overall satisfaction degree is strongly dependent on the degree of the least satisfied criterion. The peculiarity of such an operator, which also distinguishes it from the traditional "min" operator, is that the extent to which the least satisfied criterion is considered depends on its importance for the user. If it is not important at all, its satisfaction degree should not be considered, while if it is the most important criterion for the user, only its satisfaction degree is considered. This way, if we consider a document d, for which the least satisfied criterion C_k is also the least important one, the overall satisfaction degree will be greater than $C_k(d)$; it will not be C_k as it would be the case with the traditional "min" operator — the less important is the criterion, the lower its chances to represent the overall satisfaction degree.

The aggregation operator F_m is defined as follows. $F_m : [0, 1]^n \rightarrow [0, 1]$ is such that, for all document d,

$$F_m(C_1(d), \ldots, C_n(d)) = \min_{i=1,n} (\{C_i(d)\}^{\lambda_i}). \tag{4}$$

Formalizations and properties of this operator are presented in [4].

[2]If there are more than one criterion with the same priority order, the average weight and the average satisfaction degree are considered.

3. USER EVALUATION OF THE PRIORITIZED AGGREGATION OPERATORS

In [3, 4] the proposed approach for prioritized aggregation of the considered relevance dimensions has been applied to personalized IR without loss of generality. The considered personalized approach relies on four relevance dimensions: aboutness, coverage, appropriateness, and reliability. The aboutness is computed as the similarity between the document vector and the query vector. The scores of the coverage and the appropriateness criteria are computed based on a similarity of the document vector and a vector of terms representing the user profile. While the reliability represents the trust degree for a user of the source from which document comes.

3.1 Preliminary Assumptions

The prioritized aggregations approach is based on the user's indication (either explicit or implicit) of the importance order of relevance dimensions. In [3, 4] different user's behaviors have been described. In the case in which a user formulates a query with the idea of locating documents which are about the query and which also cover all his interests, and at the same time he does not care about the fact that the document also focuses on additional topics the user can be called "coverage seeker". If on the contrary the user's intent is to privilege documents which perfectly fit his interests the user is called "appropriateness seeker"

On the contrary, a user who formulates a query which has no intersection with his interests or users who do not have a defined list of interests – *interest neutral* – will not give any importance to the coverage and appropriateness criteria. Users of this kind are just looking for a satisfactory answer to their current concern, as expressed by their query. Finally, users who are cautious about the trustworthiness of the origin of the retrieved documents – *cautious* – will give more importance to the reliability criterion than to the others.

For example, *coverage seeker* users can be defined as follows:

$CARA_p$: $coverage \succ aboutness \succ reliability \succ appropriateness$;

3.2 Experiments

In this section, the impact of the proposed prioritized aggregation operators in the personalized IR setting is evaluated. In Section 3.2.1 we present the settings used to perform the experiments, while in Section 3.2.2 we discuss the obtained results.

3.2.1 Experimental Settings

The traditional way to evaluate an information retrieval system is based on a test collection composed by a document collection, a set of queries, and a set of relevance judgments which classify a document as being relevant or not for each query. Precision and recall are then computed to evaluate the effectiveness of the system. Unfortunately, there is not a test collection suited to evaluate a system based on approaches like the one proposed in this paper. It is important to notice that in the case of a user-independent aggregation of the multiple relevance numeric assessments, a traditional system's evaluation could be applied. In fact if for example the single assessment scores are aggregated by a mean operator, the system could produce the same result for a same query and a same document, independently of the user judgments. When applying the prioritized aggregation that we have proposed, a same document evaluated with respect to a same query, could produce distinct assessment scores depending on the adopted prioritized scheme, which is user-dependent.

The evaluation approach proposed in this paper is based on an analysis of how document rankings are modified accordingly to the prioritized aggregations associated with the user's typologies that we have identified in Section 3.1.

The relevance criteria and their aggregation discussed in the previous sections have been implemented on top of the well-known Apache Lucene open-source API [3]. The Reuters RCV1 Collection (over 800,000 documents) has been used. The method that we have used to generate both queries and user's profiles has been inspired by the approach presented by Sanderson in [9]. In this work the author presents a method to perform simple IR evaluations by using the Reuters collection that does not have queries nor relevance judgments, but has one or more subject codes associated with each document.

He splits the collection in two parts, a query set **"Q"** and a test set **"T"**, and documents are randomly assigned to one of the two subsets. Then, all subject codes are grouped in a set **"S"**. For each subject code s_x, all documents tagged with the subject code s_x are extracted from the set **"Q"**. From these documents, the pairs (word, weight) are generated to create a query. Then, the query is performed on the set **"T"**. The precision/recall curves are calculated by considering as relevant, the documents that contain the subject code s_x.

We have been inspired by Sanderson's approach to build both the queries and the user's profiles. The queries have been created as expressed above. The creation of the user's profile has been done in the following way. The set **"Q"** has been split in different subsets based on the subject code of each document (ex. "sport", "science", "economy", etc.). Each subset of **"Q"** represents the set of documents known by the users interested in that particular topic. For example, the subset that contains all documents tagged with the subject code "sport" represents the set of documents known by the users interested in sports.

We have indexed each subset of **"Q"** and, for each created index, we have calculated the TF-IDF of each term. Then, we have computed a normalized ranking of these terms and we have extracted the most significant ones. The TF-IDF of each term represents the interest degree of that term in the profile, that is, how much the term plays the role of a good representation of the user's interests.

An example of user's profile is illustrated in Table 1. For example, the users associated with the "BIOTECH" profile have, with respect to the term "disease", an interest degree of 0.419. Each profile is viewed as a long term information need, therefore, it is treated in the same way as documents or queries.

To study the behavior of the system, we have carried out a user evaluation as proposed in [1] [5] [6].

The user evaluation described in this paper has been inspired by the one suggested in [7] that simply consists in a procedure in which a set of at least 6 users performs a set of at least 6 queries.

In these experiments we have considered eight users with eight different profiles, each one associated with a subset of **"Q"** (Table 2).

BIOTECH					
scientist	1.000	gene	0.402	patient	0.260
researcher	0.563	study	0.386	brain	0.259
disease	0.419	clone	0.281	people	0.254
cancer	0.410	animal	0.279	experiment	0.249
human	0.406	planet	0.267	drug	0.247

Table 1: The top 15 interest terms of the BIOTECH profile.

The aims of these experiments are to verify that: (i) when a user performs queries in-line with his interests, by applying a prioritized aggregation operator, the system produces an improved ranking with respect to the one produced by simply averaging the scores, and (ii) when a user performs queries that are not-in-line with his interests, by applying a prioritized aggregation operator, the quality of the produced rank does not decrease with respect to the situation in which the prioritized aggregation operators are not applied.

Two kinds of queries have been considered. Those which are in-line with the interests contained in the user's profile, Q_i, and those which are not-in-line with the interests contained in the user's profile, Q_n. Table 2 illustrates the set Q_i and shows the associations between the user's profiles and the performed queries. In these preliminary experiments only one query has been generated for each user. For instance, for User 1, the set Q_i is composed only by the query Q1, while the set Q_n is composed by all the other queries from Q2 to Q8.

For User 2, the set Q_i is composed only by the query Q2, while the set Q_n is composed by the query Q1 and the queries from Q3 to Q8, and so on for the other users.

User	Profile Name	Query
User1	SPACE	Q1: "space shuttle missions"
User2	BIOTECH	Q2: "drug disease"
User3	HITECH	Q3: "information technology"
User4	CRIMINOLOGY	Q4: "police arrest sentence fraud"
User5	DEFENSE	Q5: "russia military navy troops"
User6	DISASTER	Q6: "flood earthquake hurricane"
User7	FASHION	Q7: "collection italian versace"
User8	SPORT	Q8: "premiership league season score"

Table 2: The queries executed for each user profile.

When a user submits a query, the matching between the query vector and each document vector is made first (aboutness), then, on each document the coverage and the appropriateness criteria are evaluated by comparing the document vector with the user's profile vector. Finally, the value of the reliability criterion, which corresponds to the degree to

[3]See URL `http://lucene.apache.org/`.

which the user trusts the source from which the document comes, is taken into account. These are the values to be aggregated — aboutness, coverage, appropriateness and reliability.

The evaluation of the produced rank is made by the eight real users that used the system. Each user analyzed the top 10 documents returned by the system and assessed, for each document, if it is relevant or not.

3.2.2 Discussion of the Results

In this section we present the obtained results. For space reasons some ranks have not been inserted, however the complete archive of the ranks produced in these experiments are available online [4]. For convenience, only the top 10 ranked documents are reported in each table. The rationale behind this decision is the fact that the majority of search result click activity (89.8%) happens on the first page of search results [11], that is, generally, users only consider the first 10 (20) documents. The baseline rank for the "Scoring" operator is obtained by applying the average operator to calculate document assessment. Such rank corresponds to the average assessment of the documents considering the four criteria and without considering priorities among the criteria. Instead, the baseline rank for the "Min" operator is obtained by applying the *standard min* operator. Table 3 illustrates an example of rank produced by the average operator after performing a query in Q_i, while Table 4 illustrates an example of rank produced by the standard min operator after performing a query in Q_i. The entries marked with the asterisk before the title, have been considered relevant with respect to both the performed query and the user profile. We can notice that there are more non-relevant documents in the top 10 list resulting from the application of average operator than in the list resulting from the application of the standard min operator. This is due to the compensatory nature of the average operator.

We illustrate the behavior of the system by taking into account different kinds of aggregations applied to the User 1, the user associated to the "SPACE" profile. In particular, we present in Tables from 5 to 10 the results obtained by applying both the Prioritized "Scoring" Operator and the Prioritized "Min" Operator, with the aggregations ACA_pR, CA_pAR, and A_pCAR

We can notice that the proposed document rankings are improved, with respect to the baselines ranking for both operators and for the considered aggregations, in the sense that the number of relevant documents in the top 10 is greater than the number of relevant documents in the baseline ranking — non relevant documents are put down in the ranking.

We can also notice that, while the document in the 9th position of the top 10 documents in Table 3 is deemed sufficiently topical for the user with profile "SPACE", the same document is not even considered in the top 10 list of any table corresponding to the prioritized "Scoring" operator. This is due to the fact that, even though the document satisfies the query because it contains information about space mission, its content is instead related to space exploration. Instead, for example, the document in the first position in the scoring baseline rank, is also proposed in almost all the top ten documents (scoring and min) including the min baseline rank. An exception is Table 6 where that document does not appear. The reason is that this document comes from a source with a very low degree of reliability.

Different considerations have to be done when the user's query is not in-line with his profile (i.e. the user's query is in the set Q_n). We will discuss about two different scenarios. In the first one the user associated with the "BIOTECH" profile executes the query associated to the "FASHION" profile, while in the second scenario, the user associated to the "CRIMINOLOGY" profile executes the query associated to the "SPACE" profile. We have noticed that, for the scoring operator, the results for all aggregations are in general similar to the baseline. The previous considerations are not valid for the prioritized min operator. It is due to its definition. Indeed, if just one criterion is weak satisfied, the overall assessment is very low. Now, if users make queries not in line with their profile, the criteria like coverage and appropriateness are weakly satisfied and then the overall value is low. Instead, when considering the prioritized min operator, the result depends also on the importance degree of the least satisfied criterion. We can conclude that the (prioritized) min operator should not be used for the users who make queries that are not in line with their profile.

[4]http://www.dti.unimi.it/dragoni/files/ MultirelevanceUserEvaluation.rar

4. CONCLUSION AND FUTURE WORK

In this paper, a user evaluation for aggregating multiple criteria has been presented and discussed.

The experimental results have been obtained thanks to a case study on personalized Information Retrieval with multi-criteria relevance. These results show that: (i) the proposed operators allow to improve the ranking of the documents which are related to the user interest, when the user formulates an interest-related query; (ii) for the "scoring" operator, when a user has no interests or formulates a query which is not related to his interests, the ranking of the documents is similar to the ranking obtaining by using the average operator; and (iii) for the "min" operator, when the user formulates a non interest-related query this operator is not suitable.

R.	Document Title	Score
1	***Shuttle Atlantis blasts off on schedule.**	0.626
2	Countdown starts for Sunday shuttle launch.	0.575
3	***Shuttle finally takes Lucid off space station Mir.**	0.573
4	U.S. spacewoman breaks another record.	0.573
5	***Shuttle Discovery heads for Florida.**	0.572
6	***Shuttle Atlantis heads for Mir despite problem.**	0.568
7	Scientists delighted with U.S. shuttle flight.	0.567
8	***U.S. shuttle launched on mission to Mir.**	0.563
9	Boeing-Lockheed group signs $7 billion shuttle pact.	0.562
10	***U.S. shuttle leaves space station Mir.**	0.561

Table 3: Results for "SPACE" profile by applying the average operator.

R.	Document Title	Score
1	***Part of planned space station arrives in Florida.**	0.250
2	***French astronaut to join Russian space mission.**	0.242
3	***Russia, hurt by Mars failure, sends probe to space.**	0.231
4	***Astronauts board shuttle for U.S. launch.**	0.228
5	***Shuttle Columbia blasts off to mission.**	0.228
6	***Shuttle Atlantis blasts off on schedule.**	0.225
7	***Shuttle Discovery lands in Florida.**	0.216
8	***U.S. space shuttle crew set for Thursday landing.**	0.215
9	***U.S. shuttle leaves space station Mir.**	0.210
10	RUSSIA: Frenchman's August Mir flight scrapped.	0.202

Table 4: Results for "SPACE" profile by applying the standard min operator.

5. REFERENCES

R.	Document Title	Score	Gap
1	***Shuttle Discovery takes off on schedule.**	1.521	25
2	***Shuttle Atlantis blasts off on schedule.**	1.427	-1
3	***U.S. space shuttle heads home.**	1.381	85
4	***Shuttle Discovery heads for Florida.**	1.333	1
5	***U.S. shuttle crew set up space laboratory.**	1.323	35
6	***Columbia shuttle mission extended one day.**	1.317	35
7	***Shuttle Atlantis heads for Mir despite problem.**	1.313	-1
8	***Shuttle Discovery lands in Florida.**	1.275	3
9	***U.S. space shuttle crew set for Thursday landing.**	1.264	62
10	***U.S. shuttle will not flush Mir's water.**	1.253	32

Table 5: Results for "SPACE" profile by applying the Prioritized Scoring Operator and ACA_pR aggregation.

R.	Document Title	Score	Gap
1	***Shuttle Atlantis to return home on Wednesday.**	0.661	53
2	***With spacewalk off, shuttle astronauts relax.**	0.652	30
3	***U.S. space shuttle heads for rendezvous with Mir.**	0.643	39
4	***U.S. shuttle crew prepares to retrieve satellite.**	0.632	257
5	***Shuttle-deployed telescope ready for action.**	0.631	260
6	***Space shuttle deploys U.S.-German satellite.**	0.628	217
7	***Shuttle crew prepares for nighttime landing.**	0.628	264
8	***Hubble service crew prepares to return home.**	0.625	150
9	***Satellites line up behind shuttle Columbia.**	0.621	129
10	RUSSIA: Sticken Mir crew stands down, says worst over.	0.620	256

Table 6: Results for "SPACE" profile by applying the Prioritized Min Operator and ACA_pR aggregation.

[1] P. Borlund. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3):152, 2003.

[2] W. S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100, 1973.

[3] C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional relevance: A new aggregation criterion. In *ECIR'09*, pages 264–275, 2009.

[4] C. da Costa Pereira, M. Dragoni, and G. Pasi. A prioritized "and" aggregation operator for multidimensional relevance assessment. In *AI*IA 2009, to appear*, 2009.

[5] P. Ingwersen. *Information Retrieval Interaction.* Taylor Graham, 1992.

[6] P. Ingwersen. Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of Documentation*, 52(1):3–50, 1996.

[7] P. Ingwersen and K. Järvelin. *The Turn Integration of Information Seeking and Retrieval in Context Series.* Springer, 2005.

[8] S. Mizzaro. Relevance: the whole history. *J. Am. Soc. Inf. Sci.*, 48(9):810–832, 1997.

[9] M. Sanderson. The reuters collection. In *Proceedings of the 16th BCS IRSG Colloquium*, 1994.

[10] T. Saracevic. The stratified model of information retrieval interaction: Extension and applications. *Journal of American Society for Information Science*, 34:313–327, 1997.

[11] A. Spink, B. Jansen, C. Blakely, and S. Koshman. A study of results overlap and uniqueness among major web search engines. *Inf. Process. Manage.*, 42(5):1379–1391, 2006.

[12] Y. C. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.*, 57(7):961–973, 2006.

R.	Document Title	Score	Gap
1	***Russians aim to fix Mir before US Shuttle arrives.**	0.777	52
2	***Russians hope to fix Mir before Shuttle arrives.**	0.742	68
3	***With spacewalk off, shuttle astronauts relax.**	0.707	53
4	Countdown continues for U.S. spacewoman's return.	0.700	70
5	***Shuttle Columbia blasts off to mission.**	0.700	137
6	***Shuttle Atlantis blasts off on schedule.**	0.682	-5
7	***Navigational problem crops up on shuttle mission.**	0.681	40
8	***U.S. shuttle launched on mission to Mir.**	0.679	0
9	Sticken Mir crew stands down, says worst over.	0.676	78
10	***Astronaut Lucid tones up for ride home.**	0.673	96

Table 7: Results for "SPACE" profile by applying the Prioritized Scoring Operator and CA_pAR aggregation.

R.	Document Title	Score	Gap
1	***Shuttle Atlantis blasts off on schedule.**	0.466	5
2	***U.S. shuttle leaves space station Mir.**	0.460	7
3	***Astronauts board shuttle for U.S. launch.**	0.459	1
4	***Shuttle Atlantis moved to pad for Mir mission.**	0.453	27
5	Russians, Ukrainian set for 1997 shuttle flights.	0.452	12
6	***Shuttle finally takes Lucid off space station Mir.**	0.450	41
7	***Shuttle Discovery takes off on schedule.**	0.447	15
8	Astronauts arrive for U.S. shuttle launch.	0.446	12
9	***U.S. shuttle launch further delayed.**	0.446	66
10	***Shuttle Columbia blasts off to mission.**	0.446	-5

Table 8: Results for "SPACE" profile by applying the Prioritized Min Operator and CA_pAR aggregation.

R.	Document Title	Score	Gap
1	***Shuttle Columbia blasts off to mission.**	0.364	141
2	***Shuttle Atlantis blasts off on schedule.**	0.364	-1
3	***Part of planned space station arrives in Florida.**	0.362	69
4	***Astronauts board shuttle for U.S. launch.**	0.351	48
5	***French astronaut to join Russian space mission.**	0.336	89
6	***Russia, hurt by Mars failure, sends probe to space.**	0.332	208
7	***U.S. shuttle leaves space station Mir.**	0.332	3
8	***U.S. space shuttle crew set for Thursday landing.**	0.314	63
9	Russians, Ukrainian set for 1997 shuttle flights.	0.303	117
10	***U.S. shuttle launched on mission to Mir.**	0.299	-2

Table 9: Results for "SPACE" profile by applying the Prioritized Scoring Operator and A_pCAR aggregation.

R.	Document Title	Score	Gap
1	***Part of planned space station arrives in Florida.**	0.250	0
2	***French astronaut to join Russian space mission.**	0.242	0
3	***Russia, hurt by Mars failure, sends probe to space.**	0.231	0
4	***Astronauts board shuttle for U.S. launch.**	0.228	0
5	***Shuttle Columbia blasts off to mission.**	0.228	0
6	***Shuttle Atlantis blasts off on schedule.**	0.225	0
7	***Shuttle Discovery lands in Florida.**	0.216	0
8	***U.S. space shuttle crew set for Thursday landing.**	0.215	0
9	***U.S. shuttle leaves space station Mir.**	0.210	0
10	Lack of funds threaten Russia's space programme.	0.204	258

Table 10: Results for "SPACE" profile by applying the Prioritized Min Operator and A_pCAR aggregation.