Multiple Similarities for Diversity in Recommender Systems

Laurent Candillier*, Max Chevalier†, Damien Dudognon*†, and Josiane Mothe†‡

* *OverBlog, Ebuzzing, Toulouse, France*
Email: firstname.name@ebuzzing.com
† *IRIT, UMR5505, CNRS, Université de Toulouse, France*
Email: firstname.name@irit.fr
‡ *IUFM, Université de Toulouse, France*
Email: firstname.name@univ-tlse2.fr

***Abstract*—Compared to search engines, recommender systems provide another means to help users to access information. Recommender systems are designed to automatically provide useful items to users. A new challenge for recommender systems is to provide diversified recommendations. In this paper, we investigate an approach to obtain more diversified recommendations using an aggregation method based on various similarity measures. This work is evaluated using three experiments: the two first ones are lab experiments and show that aggregation of various similarity measures improves accuracy and diversity. The last experiment involved real users to evaluate the aggregation method we propose. We show that this method allows the balance between accuracy and diversity of recommendations.**

Keywords-Recommender System; Diversity; Similarity Measures; Users Study; Information Retrieval

I. INTRODUCTION

As explained by Ricci *et al.* [2], "Recommender Systems (RS) are software tools and techniques providing suggestions for items to be of use to a user". RS are uselly classified according to how item suggestions are generated; three categories are genearlly distinguished [3], [4]:

- Collaborative filtering that uses social knowledge to generate recommendations;
- Content-based filtering that uses content features to generate recommendations;
- And hybrid filtering that mixes content-based and collaborative filtering approaches.

Kumar and Thambidurai underline in [5] that "recommender systems are characterized by cross-fertilization of various research fields such as: Information Retrieval, Artificial Intelligence, Knowledge Representation, Discovery and Data/Text Mining, Computational Learning and Intelligent and Adaptive Agents". Indeed, when considering content-based filtering techniques, an important issue is to match various items and to identify those that should be recommend to a given user. In content-based R, such a matching is mainly based on similarity measures coming from Information Retrieval (IR) field [6].

IR usually sorts the retrieved documents according to their similarity with the user's query [7]. Doing so, IR systems assume that document relevance can be calculated independently from other documents [8]. As opposed to this assumption, various studies consider a user may prefer to get documents treating of various aspects of her information need rather than possibly redundant aspects within documents [9], [10]. Diversity pursues this goal.

Document diversity has many applications in IR. First, it is considered to be one solution to query term ambiguity [8]. Indeed, queries as expressed by users are not enough to disambiguate terms. To answer ambiguous queries, Clarke *et al.* [8] suggest that IR can provide the user with a range of documents that corresponds to the various term senses. In that case, redundancy can be penalized by lowering the rank of a document that is too similar to a document ranked higher in the list. Following the same idea and to face query term ambiguity, Chifu and Ionescu propose a non-supervised method that clusters documents and re-order retrieved documents based on the clustering results [11].

Diversity became a real challenge in RS field too [12]. It aims at tackling at least two objectives: removing redundancy in the recommendation list (i.e. avoiding recommendation of items that are too similar) and taking into account diverse interests.

In the literature two kinds of diversity have been proposed: individual diversity and aggregate diversity [13]. Individual diversity aims at recommending to a single user some recommendations that are not redundant; aggregate diversity aims at recommending items that are not redundant from one user to another (considering the "long tail" phenomenon). This paper focuses on individual diversity to provide a user with a diversified list of recommendations.

In order to achieve this goal, we investigate the relation between diversity and similarity measures. We study how different similarity measures, based on various aspects of recommended items, can be aggregated to provide more diversified recommendations while keeping a good accuracy.

Indeed, our main objective being to consider the variety of the users' expectations, the recommended items must be sufficiently diversified to cover a large range of expectations. This intuition comes from the fact that item relevance may

be multi-dimensional and dynamic [14]. This idea was initially developed by Candillier *et al.* [1] and is extended in this paper.

The paper is organized as follows: Section II presents the related works dealing with the links between similarity measures and diversity. We describe in Section III two first experiments based on TREC [15] IR tasks (*adhoc* and *diversity*). These experiments show that aggregation of various similarity measures may improve accuracy and diversity. In Section IV, we complete these experiments with a user study on a blog platform consisting of more than 20 million of articles. We show the positive impact of the aggregation of various similarity measures on the users' perception of diversity in recommendations. Section V concludes this paper and underlines our future work.

II. Related works

In this section, we explain that diversity can result from the use of various similarity measures (notice that similarity measures used in RS mostly come from IR).

Users' interests are different, multidimensional and dynamic [14]. This assumption is confirmed forasmuch as document usefulness can be estimated differently. Mothe and Sahut [16] consider that a document can be evaluated on various criteria:

- Relevance;
- Information validity;
- Physical and ergonomic aspects.

Each of these criterium being in turn depicted by several sub criteria.

To deal with the variety of interests, IR systems diversify the retrieved documents [17], [12]. Doing this, the systems maximize the chances of retrieving at least one relevant document to the user [18].

IR literature distinguishes topicality and topical diversity. Topicality makes reference to which extent the document may be related to a particular topic [19] and is not related to diversity. Topical diversity refers both to extrinsic diversity and intrinsic diversity. The former helps to dispel the uncertainty resulting from the ambiguity of the user's needs or from the lack of knowledge about user's needs [20]. The intrinsic diversity, or novelty, intends to eliminate redundancy in the retrieved documents [8]. Very similar documents allow the system to increase the accuracy but do not improve the user's satisfaction [21]. The intrinsic diversity allows the system to present to the user:

- Various points of view;
- An overview of the topic that can only be achieved by considering simultaneously several documents;
- Or even to check the information reliability [20].

Topical diversity is generally used to reorder the retrieved documents. Two types of methods are generally used. The first one considers the reordering process as a clustering problem, while the other is based on a selection method such as the Maximal Marginal Relevance (MMR) proposed in [9].

With regard to clustering method, He *et al.* [22] use Single Pass Clustering (SPC). In this approach, the first document in the result list is selected and assigned to the first cluster. Then, the algorithm processes down the list of retrieved documents and assigned each document to the nearest cluster. If the document-cluster similarity is below a defined threshold, the document is assigned to a new cluster. Bi *et al.* [23] obtained better results using the k-means algorithm [24]. Whatever the algorithm used, assignment to different clusters is generally done using a distance such as the Euclidean distance or the Cosine measure, eventually weighted by the terms frequency. Meij *et al.* [22] apply a hierarchical clustering algorithm on the top fifty retrieved documents using document language modeling based approach. The document selection phase used to build the result list is based on cluster quality and stability metrics. Then, the best documents from each cluster are selected.

In these approaches, the clustering step takes place after a set of documents has been retrieved; the documents are grouped together according to the sub-topics clusters identify.

Topical diversity is also used to reduce redundancy in the retrieved document list. MMR [9] or sliding window approaches [25] aim at selecting the documents maximizing the similarity with the query and, at the same time, minimizing the similarity with all the documents already selected. The function used to compute the similarity between a given document and the documents already selected can differ from the similarity function used to estimate the relevance with the query [9].

Several approaches select the documents to be reordered using indicators or filters to increase the diversity in the results. Kaptein *et al.* [26] employ two types of document filters: a filter, which considers the number of new terms brought by the document to the current results and a link filter, which uses the value added by new input or output links to select new documents. Furthermore, Ziegler *et al.* [21] propose an intra-list similarity metric to estimate the diversity of the recommended list. This metric uses a taxonomy-based classification.

However, some user's needs cannot be simply satisfied by topic-related documents. For instance, serendipity aims at bringing to the user attractive and surprising documents she might not have otherwise discovered [27]. It is an alternative to topical diversity. For example Lathia *et al.* [28] investigate the case of temporal diversity. In the other hand, Cabanac *et al.* [29] consider organizational similarity that considers how the users sort their documents in a folder hierarchy.

Thus, similarity measures are different and may either be based on document content or structure, or on document usage considering popularity or collaborative search.

In the literature, several types of similarity functions have been considered:

- Based on document content: to be similar two documents should share indexing terms. Example of such measures are the Cosine measure [7], or semantic measures [30], [31];
- Based on document popularity such as the BlogRank [26];
- Collaborative: the document score depends on the scores that previous users assigned to it [32];
- Based on browsing and classification: document similarity is either based on browsing path [33] or considering the categories users assigned to viewed documents [23];
- Based on relationships: social similarity functions use relationships between contents and users [34], [35].

In this context, we hypothetise that diversification of recommendations can be obtained by combining several similarity metrics. The reason is that each metric answers a specific need or represents a particular view of the information interest. Similarly, Ben Jabeur *et al.* [34] combined a content similarity measure, based on TF-IDF [36], with a social measure which reflects the relationships in a social network of authors. The main difficulty with this kind of approaches lies in the way of combining the similarity measures. Whether it is a linear combination, or a successive application of measures, a combination boils down to give some importance to each measure and to favor certain facets over others.

An alternative to similarity combination is to consider different similarity metrics independently. Amazon.com [37] offers several recommendation lists to the user and indicates the type of measure used in naming these lists (e.g. "Customers who viewed this item also bought", "Inspired by the general trends of your purchases"). However, this independence of similarity metrics sometimes leads to a redundancy of information: one document can be recommended to the user in several lists of recommendations.

Fusion approaches offer a way to solve this problem. Indeed, the fusion of results from different similarity metrics within a single list of recommendations eliminates duplicates. Shafer *et al.* [18] and Jahrer *et al.* [38] propose to merge multiple sources of recommendations and therefore present a "Meta RS". Depending on the fusion approach, it is possible to favor documents appearing in multiple lists or not [39].

Finally, a graph approach can be used to fuse a set of similarity measures [40]. The results of each measure help to establish links between documents. These links are materialized by edges in a graph, weighted by the similarity scores and the documents are represented by nodes. The number of edges between two documents is only limited by the number of similarity measures used.

To be able to evaluate and compare topical diversity oriented approaches, TREC Web 2009 campaign [15] defines a dedicated topical diversity task. This task is based on the ClueWeb09 dataset, which consists in roughly one billion web pages crawled during January and February 2009. The size of the whole dataset (named Category A) is about twenty five Terabytes of data in multiple languages. The set B of the corpus we use for our experiments only focuses on a subset of English-language documents, roughly fifty million documents. The *diversity* task uses the same fifty queries as the *adhoc* tasks [41].

Clarke *et al.* [42] present the panel of metrics used to estimate and compare the performances of the topical diversity approaches. In our experiments, we only consider the Normalized Discounted Cumulative Gain (α-nDCG) [8] which is the metric used for the TREC Web 2009 evaluation campaign.

This evaluation framework is not enough to evaluate RS diversity when not only content-based elements are used but others also. Indeed, it turns out that the proposed approaches, either based on clustering algorithms or on selection criteria, are mainly focused on content and on topical diversity. The available evaluation frameworks, such as the TREC Web *diversity* task, have been designed to measure the performances of these content-based approaches. To be able to evaluate other types of diversity, like serendipity, and to truly gauge the user's satisfaction, a user study is necessary [43].

The hypothesis of our work is that diversity obtained when aggregating the lists of items resulting from different similarity measures is a means to diversify recommendations in a RS. Indeed, even if a unique recommendation method is efficient in the majority of the cases, it is useful to consider other users' expectations. Content-based diversity, but also other sorts of diversity, should be considered in recommendations.

This paper aims at showing the impact on diversity in RS of an aggregation method applied to various similarity measures. To achieve this goal we propose to verify three hypotheses:

- The aggregation of similarity measures considering the same aspect of item (e.g. item topic) improves the accuracy of recommended items;
- The aggregation of a variety of similarity measures improves the overall diversity of recommended items;
- The users' perception of diversity is high when aggregating various similarity measures while keeping a perception of a good level of accuracy.

These hypotheses are studied in this paper through three experiments we conducted.

III. Experiments

We hypothesize there is not one single approach that can satisfy the various users' expectations, but a set of

complementary approaches. In our view, each approach could correspond to a different point of view on the information and thus answers to specific users' expectations. We hypothesize that aggregating various approaches could be a relevant solution. To start with, we decided to verify that two distinct approaches retrieve different documents for a given IR tasks (*adhoc*, *diversity*). We then show the positive impact of the aggregation of these distinct approaches on accuracy.

For the experiment, we consider several systems, which were evaluated within the same framework to ensure they are comparable, and for which the evaluation runs were available. We focus on the *adhoc* and *diversity* tasks of the TREC Web 2009 campaign considering only the set B of the corpus to get comparable systems. Moreover, we choose the four best runs for each task rather than taking into account all the submitted ones.

To compare the selected runs, we follow the framework and the metric proposed by Lee [44] in the context of IR. This framework is widely used in the literature. We compute the overlap for each pair of runs, that is to say the number of common documents between the two compared runs. The overlap is computed for the n first documents. We first compare the global overlap considering all retrieved documents. Then, we focus on the relevant document overlap and on the non relevant document overlap.

We use the metric proposed by Lee [44] and defined as follows:

$$overlap = \frac{2 \cdot |run_1 \cap run_2|}{|run_1| + |run_2|} \quad (1)$$

Where run_1 and run_2 are the documents of the two runs to be compared. The value of the overlap is between 0, when both runs have no common document, and 1 if the same documents are retrieved by run_1 and run_2.

In this section, we compare the results obtained by the best runs in two tasks: *adhoc* task and *diversity* task.

A. Adhoc task experiment

1) Adhoc task and compared runs: The TREC *adhoc* task is designed to evaluate the performances of systems, that is to say their ability to retrieve relevant documents for a given query. These systems have to return a ranked list of documents from the collection, ordered by decreasing expected relevance. The expected relevance considers each document independently: it does not take into account the other documents that appear before it in the retrieved list. The full evaluation counts fifty queries [41].

The performances of the different evaluated systems are compared using MAP which is based on precision. The precision P defines the proportion of relevant documents among the retrieved documents and is formally expressed by:

$$P = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad (2)$$

Thus, the average precision $AveP$ is defined as:

$$AveP = \frac{\sum_{k=1}^{n} (P(k) \cdot rel(k))}{|\{relevant\ documents\}|} \quad (3)$$

Where

- $P(k)$ is the precision considering the k first documents in the result list;
- $rel(k)$ is a function which indicates if a document is relevant (1) or not (0).

Finally, the MAP measure, used for the TREC evaluation campaigns, is the mean of the average precision scores $AveP$ for each query q of the set of queries Q:

$$MAP = \frac{\sum_{q=1}^{|Q|} AveP(q)}{|Q|} \quad (4)$$

The scores of the four best run at the TREC Web 2009 *adhoc* task are presented in Table I.

Table I
TREC WEB 2009 ADHOC TASK RESULTS

Group Id	Run Id	MAP
UDel	udelIndDRSP	0.2202
UMD	UMHOOsd	0.2142
uogTr	uogTrdphCEwP	0.2072
EceUdel	UDWAxQEWeb	0.1999

The runs (Run Id) we kept are the following ones:

- udelIndDRSP: this run, generated using the Indri search engine [45], combines the query-likelihood language model with the Markov Random Fields (MRF) model of term dependencies and the pseudo relevance feedback with relevance models. It also uses a metric to define trust in a domain. This metric is supported by content filtering whitelists and blacklists and publicly-available sendmail [46];
- UDWaxQEWeb: relies on an axiomatic retrieval framework where the relevance is modeled with retrieval constraints. The retrieval process aims at searching for functions that can satisfy these constraints. The approach is completed by a query expansion step. The expansion terms semantically related to the query terms are extracted from a Web search engine [47];
- UMHOOsd: uses a model based on the MRF in a distributed retrieval system built on Hadoop [48], the open source implementation of MapReduce [49];
- uogTrdphCEwP: uses the Terrier IR platform [50] with the implementation of the DPH weighting model derived from the Divergence From Randomness (DFR)

model. A query expansion step completes the retrieval process using the ClueWeb09 Wikipedia [51] documents [52].

2) Results:

a) Overlap of retrieved documents: Figure 1 presents the average overlap and precision for the four runs selected in the *adhoc* experiment, considering the fifty queries of the task. The precision and the overlap both take their values in between 0 and 1. We note that when we focus only on the first retrieved documents the global overlap is low, in spite of the fact that the first retrieved documents are most relevant. For example, taking the ten first documents for which the precision reaches its highest value (0.386), the average global overlap is only 0.255. The global overlap is low, even on a set of hundred documents (0.390).

Table II
GLOBAL OVERLAP CONSIDERING THE RUNS OF THE TREC WEB 2009 ADHOC TASK

Runs		udwa	umhoo	udel
umhoo	Relevant	0.8120		
	Non Relevant	0.5958		
udel	Relevant	0.7616	0.7806	
	Non Relevant	0.4721	0.5177	
uog	Relevant	0.7223	0.7583	0.6915
	Non Relevant	0.5133	0.4754	0.4066
Average	Relevant		0.7544	
	Non Relevant		0.4968	

Next, we focus on the average global overlap of relevant and non-relevant documents. We first compute the overlap (see Table II) for the overall runs, that is to say considering one thousand documents per query. We obtain an average global overlap equals to 0.754 for the relevant documents, and 0.497 for the non-relevant ones. These results are consistent with Lee's conclusions [44] on the TREC3 *adhoc* task: different runs retrieve different non-relevant documents but retrieve similar relevant documents.

Generally speaking, IR users focus on the first documents only [53]. In the same way, in the context of RS, only a small set of recommendations is proposed to the user. The choice is harder when there are a lot of documents provided to the user [54].

Therefore, we further analyze the evolution of relevant and non-relevant document overlap depending on the number of retrieved documents. Figure 2 shows that when we consider at the fifty top documents, the overlap is low for both relevant and non-relevant documents and it is pretty much the same until twenty documents.

b) Aggregating retrieved documents: The experiment demonstrates that, for a given query, two distinct systems are unlikely to retrieve the same set of top documents. Therefore, it is reasonable to expect that system result aggregation is relevant and could help to improve the accuracy of the

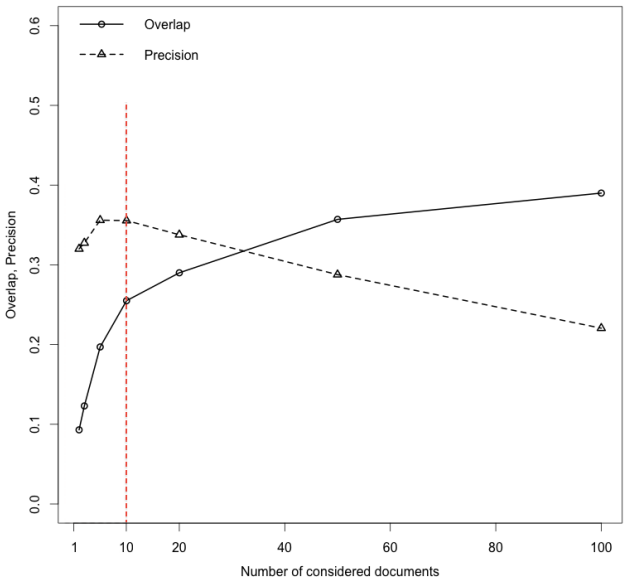

Figure 1. Average global overlap and precision for TREC Web 2009 adhoc task

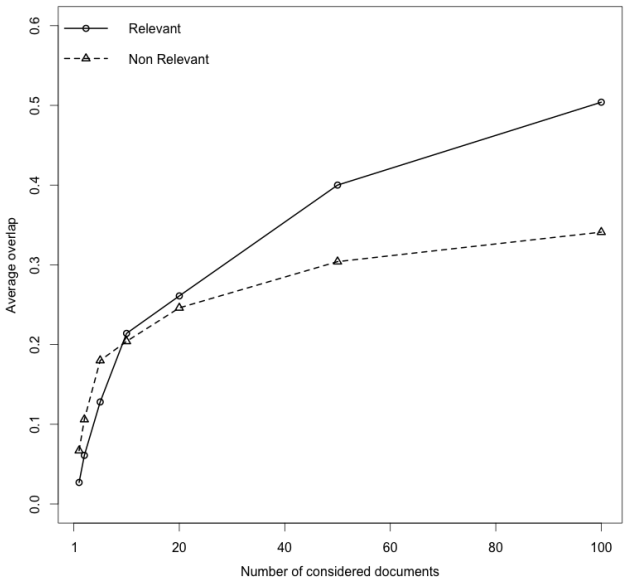

Figure 2. Average overlap for TREC Web 2009 adhoc task considering relevant and non relevant documents

results. To assess the relevance of approach aggregation, we aggregate the four runs previously used. For each query, all the retrieved document sets are aggregated using the fusion CombMNZ function [55] to generate a new run. CombMNZ has shown to be very efficient in the IR context.

$$CombMNZ(d_i) = (\sum_{j=1}^{n} w_{ij}) \cdot Count(d_i) \qquad (5)$$

Where

- d_i is a document;
- n is the number of similarity measures ;
- w_{ij} is the document score obtained with a similarity measure;
- $Count$ is a function which indicates the number of similarity measures that have retrieved the document d_i.

Then, following the *adhoc* task evaluation framework, we compute the MAP and obtain a score of 0.237, which outperforms the best run (0.2202).

In the next experiment, we apply the framework used on the *adhoc* task to the *diversity* task. We first aim at checking if various diversity-oriented approaches retrieve the same relevant documents. Subsequently, we investigate the consequences of approach aggregation on the diversity in the retrieved documents.

B. Diversity task experiment

1) Diversity task and compared runs: Similarly to the previous experiment, we center on several systems submitted at the TREC Web *diversity* task. All these systems aim at providing users with diversified result lists. The goal of the *diversity* task is to retrieve a ranked set of documents that together provide complete coverage for a query. Moreover, excessive redundancy should be avoided in the result list. The probability of relevance of a document depends on the documents that appear before it in the result list [41]. The queries are the same for the *adhoc* and the *diversity* tasks. The evaluation measures and the judging process differ from the *adhoc* task. The measure used for the TREC Web 2009 *diversity* task is the α-*nDCG* [8] derived from the Discounted Cumulative Gain (DCG) proposed in [25].

The DCG is based on the gain vector G and on the Cumulative Gain CG defined as:

$$G[k] = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}} \quad (6)$$

$$CG[k] = \sum_{j=1}^{k} G[j] \quad (7)$$

Where $J(d_k, i)$ is equal to 1 if the kth document is judged as relevant and 0 otherwise. Thus, DCG is formalized by:

$$DCG[k] = \sum_{j=1}^{k} \frac{G[j]}{log_2(1+j)} \quad (8)$$

Normalized Discounted Cumulative Gain ($nDCG$) is the ratio between the Discounted Cumulative Gain DCG and the ideal Discounted Cumulative Gain DCG':

$$nDCG[k] = \frac{DCG[k]}{DCG'[k]} \quad (9)$$

For the evaluation process, α is set to 0.5 according to [8]. Table III presents the scores obtained by the different systems at their best run.

Table III
TREC WEB 2009 DIVERSITY TASK RESULTS

Group Id	Run Id	α-nDCG@10
Waterloo	Uwgym	0.369
uogTr	uogTrDYCcsB	0.282
ICTNET	ICTNETDivR3	0.272
Amsterdam	UamsDancTFb1	0.257

For the *diversity* task, we retained the following runs:

- uwgym: this run acts as a baseline run for the track and should not be considered as an official run. It was generated by submitting the queries to one of the major commercial search engines. The results were filtered to keep only the documents included in the set B of the ClueWeb collection [41];
- uogTrDyCcsB: similarly to the *adhoc* task, this runs relies upon the DPH DFR weighting model but uses a cluster-based query expansion technique, using the Wikipedia documents retrieved [52];
- ICTNETDivR3: this run applies the k-means clustering algorithm to the set of documents retrieved at the *adhoc* task. A document is assigned to the nearest cluster using Euclidean distance or Cosine measure. Each cluster identified represents a subtopic of the query [23];
- UamsDancTFb1: this run uses a sliding window approach that intends to maximize the similarity with the query and, at the same time, to minimize the similarity with the previous selected document. The documents are selected depending on two metrics: Term Filter (TF) and Link Filter (LF). TF focuses on the number of new unique terms to select a new document, while LF uses the new incoming or outgoing links. The document bringing the most new information (links or terms) is selected [26].

2) Results:

a) Overlap of retrieved document sets: As shown in Figure 3, the behavior observed in the previous experiment is more pronounced: the global overlap does not exceed 0.1, even when one hundred retrieved document lists are considered.

These observations are also true when we focus only on relevant and non-relevant documents (see Figure 4), independently of the number of documents considered. In fact, Table IV shows the overlap reaches 0.238 for relevant documents and 0.065 for non-relevant documents when the overall runs (thousand documents) are taken into account. These results confirm our hypothesis that distinct approaches produce distinct results, even if they attempt to reach the same goal.

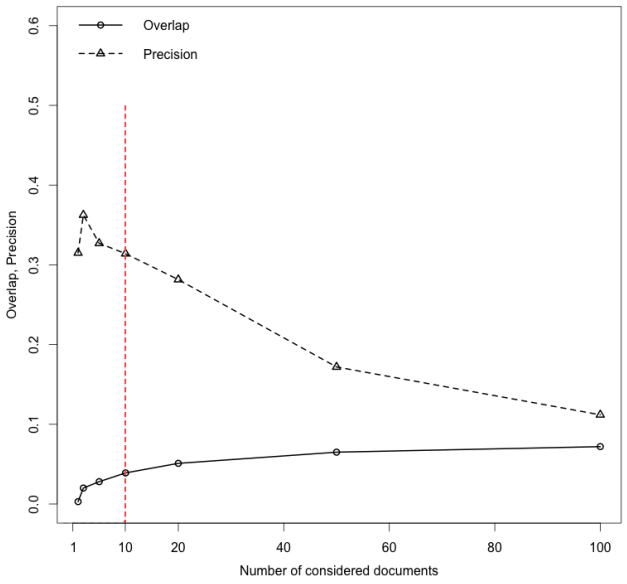

Figure 3. Average global overlap and precision for TREC Web 2009 diversity task

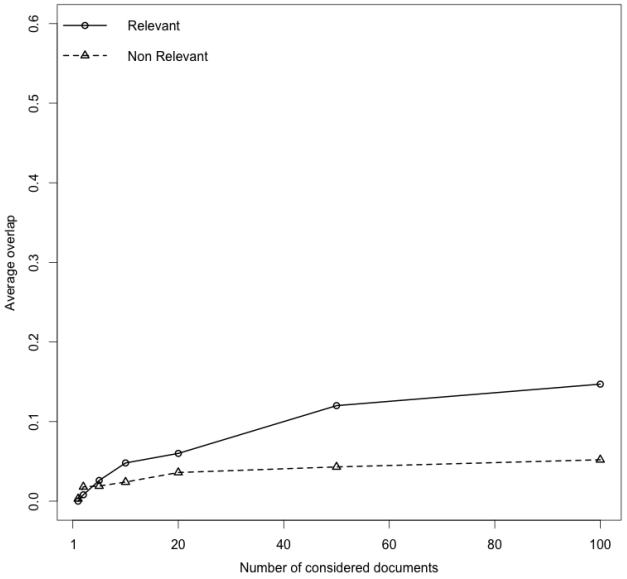

Figure 4. Average overlap for TREC Web 2009 diversity task considering relevant and non relevant documents

b) Aggregating retrieved documents: In the same manner as for the *adhoc* task experiment, we aggregate the runs to check if it helps to bring more *diversity* in the retrieved documents. However we do not use uwgym run which should not be considered as an official run [41]. The aggregation step is also based on the *CombMNZ* function. Finally, we compute α-*nDCG* for the generated run and we obtain 0.283. Although this score stays below the score of the uwgym run which acts as the baseline, it outperforms the best official run (0.282). It confirms that aggregating such approaches produce a more diversified list.

Table IV
GLOBAL OVERLAP FOR THE OVERALL RUNS OF THE TREC WEB 2009 DIVERSITY TASK

Runs	Documents considered	ictnet	uams	uog
uams	Relevant	0.1953		
	Non Relevant	0.0456		
uog	Relevant	0.4051	0.2823	
	Non Relevant	0.1818	0.1052	
uwgym	Relevant	0.1498	0.2095	0.1870
	Non Relevant	0.0154	0.0217	0.0177
Average	Relevant		0.2382	
	Non Relevant		0.0645	

C. Conclusion on the impact of the aggregation method

Whatever is the purpose of the different approaches, whether they intend to diversify the recommended items or whether they are designed to retrieve items matching the users' needs (e.g. topical search), the overlap between the lists of items they retrieve is low. Few documents are retrieved in multiple lists. We note that this observation is especially true when we consider only the first documents, which should theoretically be the most relevant. Finally, the experiment demonstrates that the aggregation of results coming from the selected systems improves accuracy and diversity.

The last experiment we present in Section IV is designed to evaluate the users' perception of diversity and accuracy of a recommendations resulting from the aggregation of various similarity measures. This experiment is conducted thanks to a RS we integrate in a blog platform.

IV. USERS STUDY: THE CASE OF OVERBLOG

A. Diversifying recommendations

We conducted a user experiment to check hypotheses about the relevance of providing diversified recommendations to users in RS while keeping a good level of accuracy. The hypotheses are:

- Most of the time, IR users search for focus information (topicality);
- Sometimes, users want to enlarge the subject they are interested in (topical diversity);
- Some users are in a process of discovering and searching for new information (serendipity);
- The interesting links between documents do not only concern the similarity of their content;
- The integration of diversity in a RS process is valuable because it allows the system to answer additional users' needs.

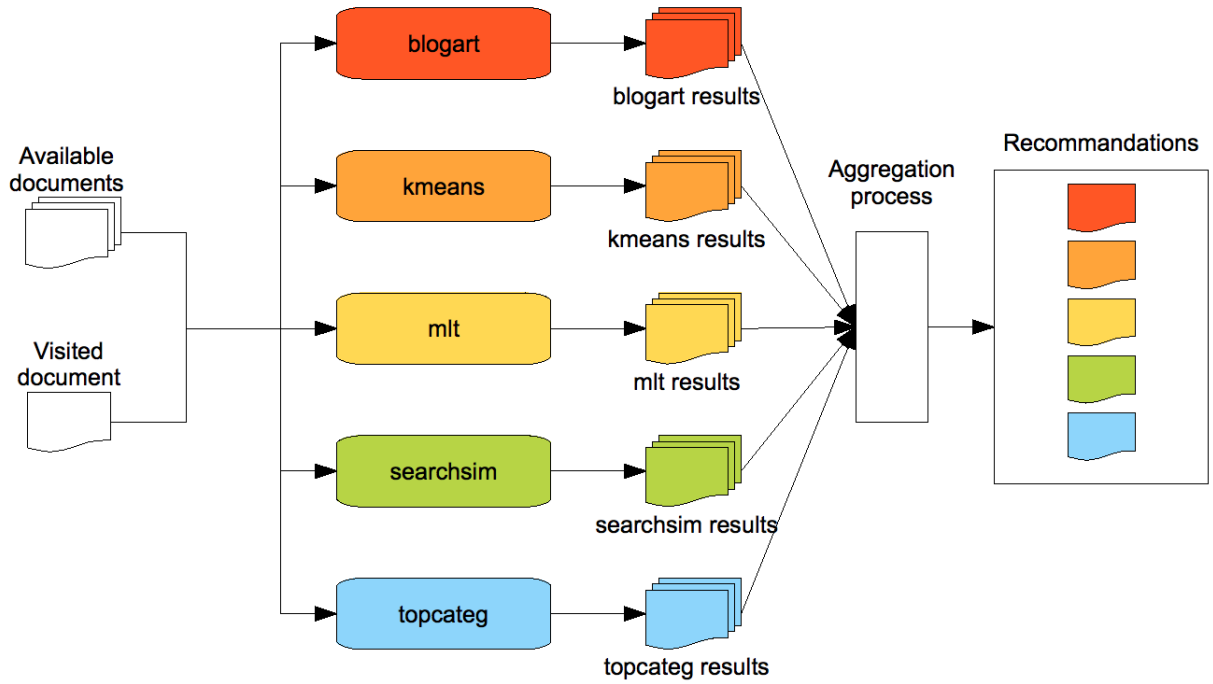

Figure 5. OverBlog aggregation prototype architecture

To check these hypotheses, we recruited 34 Master students in management, fluent in French, and asked them to test and compare various RS. This task lasts about one hour. The users were first asked to type a query on our search engine (first time the query was set, to ensure overlap about the documents they all considered, and then a query of their choice). They had to choose one relevant document and were then shown two lists of recommended documents related to this document:

- One list was based on one of the five similarity measures we designed: *mlt* and *searchsim* that use topicality, *kmeans* that uses topical diversity, and *topcateg* or *blogart* that use serendipity (see system description Section IV-B). These measures act as baselines;
- The other list was our RS, designed by aggregating the results of those five previous defined similarity measures (choosing the first document in the result list for each measure).

Each resulting list contained five documents, and the users did not know which measure it corresponds to. They were then asked to choose which list they found the most relevant, and which one they found the most diversified.

Finally, the two lists were mixed into one, and the users had to assess which documents were relevant according to them.

B. Data and systems

In this experiment, we focused on the French documents of the OverBlog platform [56]. The data used represent more than twenty million articles distributed on more than one million blogs.

We use five similarity measures that have been applied to OverBlog documents to get various recommendation lists which are then aggregated. We define a similarity measure as a function which associates a initial document d_0 (the document visited by the user) with a set of couple (d_i, w_i) where d_i is a document from the available collection and w_i the weight affected to this document by the function. It can be formalized as follows:

$$f(d_0) = \{(d_i, w_i)\} \tag{10}$$

The OverBlog similarity measures are:

- *blogart* (serendipity): returns documents randomly selected in the same blog of the visited document. The author is also the same;
- *kmeans* (topical diversity): classifies the documents retrieved by the Solr search engine [57] with the k-means clustering algorithm [24]. The documents retrieved are those that are most similar to the title of the visited document. We assume each cluster corresponds to one sub-topic of the document subject. The final result list is built by picking up in each cluster the document with the higher score in the initial result list;

- *mlt* (topicality): uses Solr MoreLikeThis module to retrieve similar documents considering all the content of the visited document. The MoreLikeThis module extracts ten representative terms within the visited document. These terms are chosen according to their frequency in the overall corpus and in the document, and then are used to selected similar documents;
- *searchsim* (topicality): uses the Solr search engine which is based on a vector-space model to retrieved documents similar to the title of the visited document;
- *topcateg* (serendipity): retrieves the most popular documents randomly selected in the same category (e.g. "High-tech", "Entertainment", ...) from the OverBlog hierarchy as the visited document. The number of its unique visitors defines the document popularity the day before.

Figure 5 presents the prototype architecture we use to recommend blog articles during the users study. According to this architecture, the available collection and the visited document, each similarity measure independently retrieves an ordered set of documents. These results constitute input data for the aggregation process that picks up the best document from each system. The final recommendation list counts five distinct documents, one per similarity measure.

The use of these five measures aims at simulating the various types of diversity (topicality, topical diversity, serendipity) and intents to limit the overlap between the documents they retrieve.

To ensure that the similarity measures used in the user study retrieve distinct results, we compute the overlap between each pair, similarly to the previous experiments described in Section III.

We observe in Figure 6 the same trends as in the experiments led on the *adhoc* and *diversity* tasks: the overlap is low between the similarity measures based on content similarities (*mlt*, *searchsim* and *kmeans*) and is null in the case of serendipity (*blogart*, *topcateg*).

C. Results

Table V shows the feedback the user panel gave concerning the interest of the proposed lists, and their feeling on the document diversity. For example (4th row), 76.5% of the lists provided by *mlt* measure have been considered as more relevant than the aggregated lists. We can see that the similarity measures perceived as the most relevant are those that focus on topicality.

The aggregated recommendations are seen as more relevant than recommendations coming from other similarity measures roughly once upon two times on average. We get the same result for *blogart* similarity measure. This is more surprising, but confirms that users' expectations sometimes do not concern the document content only.

The answers to the question "Which one of the following result lists seems the most diversified to you?" are even more surprising: there are not high differences between the systems, and the aggregated system is seen, on average, as more diverse in 50% of cases. We think this might be explained by the fact that users have difficulties in defining the notion of diversity. We should have probably helped them by clarifying the question we asked.

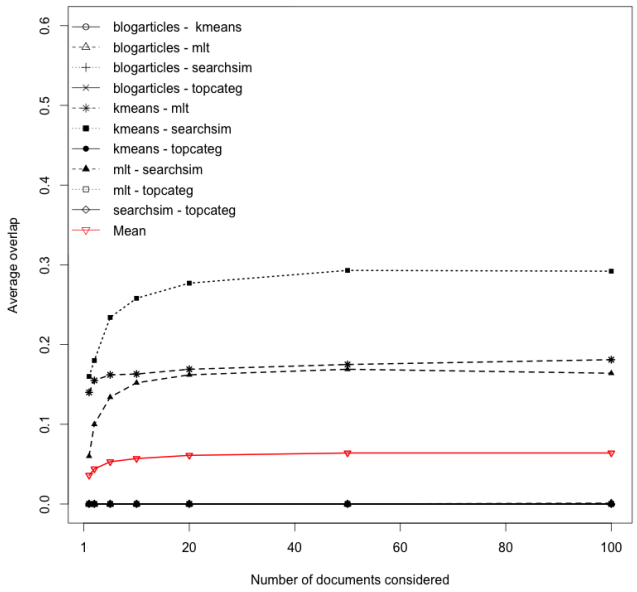

Figure 6. Average overlap for the different results obtained by the five OverBlog similarity

Table V
PERCENTAGE OF USERS WHO CONSIDER THE SYSTEM TO BE MORE RELEVANT/DIVERSIFIED THAN THE AGGREGATED SYSTEM

System	Relevance	Diversity
blogart	44.7%	55.3%
kmeans	70.8%	33.3%
mlt	76.5%	50.0%
searchsim	64.3%	42.9%
topcateg	15.4%	65.4%

Table VI describes the precision of each similarity measure, that is to say the proportion of relevant documents within the retrieved document set. Results confirm the approaches that use content similarities are seen as more relevant. *kmeans*, that proposes topical diversity, has the best results. On the contrary, *topcateg* and *blogart* that search for serendipity have lower results.

As expected, the aggregated recommendations offer an interesting compromise between these different similarity measures and a good balance between diversity (previous result) and precision. Indeed, it obtains a precision value of 0.267 that is higher than the average precision of other similarity measures (0.228). Even if it is lower than the best one (*kmeans*), this result is encouraging regarding the very

Table VI
PRECISION PER SYSTEM

System	*blogart*	*kmeans*	*mlt*	*searchsim*	*topcateg*	*aggregated*
Precision	0.147	0.385	0.265	0.307	0.038	0.267

Table VII
DISTRIBUTION OF THE RELEVANT DOCUMENTS

System: *aggregated* against	*blogart*	*kmeans*	*mlt*	*searchsim*	*topcateg*
Retrieved by the system only	35.00%	52.46%	54.69%	52.43%	8.77%
Retrieved by aggregated system only	65.00%	21.31%	32.81%	38.83%	91.23%
Commons	0.00%	26.23%	12.50%	8.74%	0.00%

low precision value of *topcateg*, *blogart* measures. In fact the low precision value of those measures may introduce noise in the recommendations, which consequently affects the overall precision. At the same time, this loss in precision is not surprising since the result of the aggregation is more diversified: it is considered as more diversified in more than 50% of cases on average. Such negative effect of diversity on accuracy has already been illustrated in [13].

Finally, Table VII compares the aggregated system with the others. It gives the proportion of relevant documents that have been retrieved by each similarity measure. For example, when comparing *mlt* to *aggregated* (4th column), 54.69% of the relevant documents have been retrieved by *mlt* only, 32.81% by *aggregated* only and 12.50% by both only. We can thus observe that, even if more relevant documents come from the similarity measures searching for topicality, a significant part of them comes from the *aggregated* system. Compared to the first experiment (Section III), we think that this result justifies our approach, because more than 20% of relevant documents are retrieved by our system only. It means that one document among the five that are proposed is considered as relevant and would not have been returned when using any system alone.

D. Users study conclusion

The *aggregated* system we propose offers a new framework to combine various similarity measures to recommend items to users. The one implemented and tested here does not outperform the others, but that was not our goal. Rather, our idea is to promote diversity, and we have seen with the user experiments that this is a relevant track. Indeed, by diversifying our recommendations, we are able to answer different and additional users' needs, when the other similarity measures focus on the majority needs: most often the content similarity. The measures we tested for serendipity were quite simple. Nevertheless, the results they returned were considered as relevant by users, and we think this is an encouraging result for improving RS since users are interested in various forms of diversity in result lists.

V. CONCLUSIONS

Users have different expectations when searching and browsing information. Systems that aim at providing tailored results to users should consider this fact. IR systems and RS should aim at answering various facets of the information needs, especially since users become used to be given personalized tools. Diversity in system answers is a way to answer this issue.

In this paper, we have shown the impact of the aggregation of various similarity measures on recommendation diversity.

Our first contribution has been to study the overlap between the documents retrieved by several IR approaches from the literature using the TREC Web 2009 datasets (*adhoc* and *diversity*) and the impact of the aggregation approach on accuracy and diversity. For the *adhoc* task, we have demonstrated that different approaches retrieve different relevant documents even if based on the same aspect of documents as the document topic. The average overlap of the result lists is low, even when the first hundred documents are considered. Moreover, this experiment has underlined an improvement of the accuracy inferred when aggregation is applied. We have also investigated the overlap for topical diversity oriented approaches and obtained similar conclusions: two distinct approaches are unlikely to retrieve the same relevant documents. In the context of topical diversity, we have proved the positive impact of the aggregation approach on recommendation diversity.

Although those approaches are all topical similarity-based, we have noted that they are based on different underlying assumptions, which explains that their overlap is low. The low overlap between the relevant retrieved documents indicates that a perfect system which would be able to satisfy the diversity of the users' needs does not exist, but rather a set of complementary approaches does. This result was the main argument in favor to the approach we defined which aims at aggregating various recommanded item lists.

To validate our proposition in a real context, we conducted a users study. This study aimed at checking if the aggregation of various similarity measures based on topicality

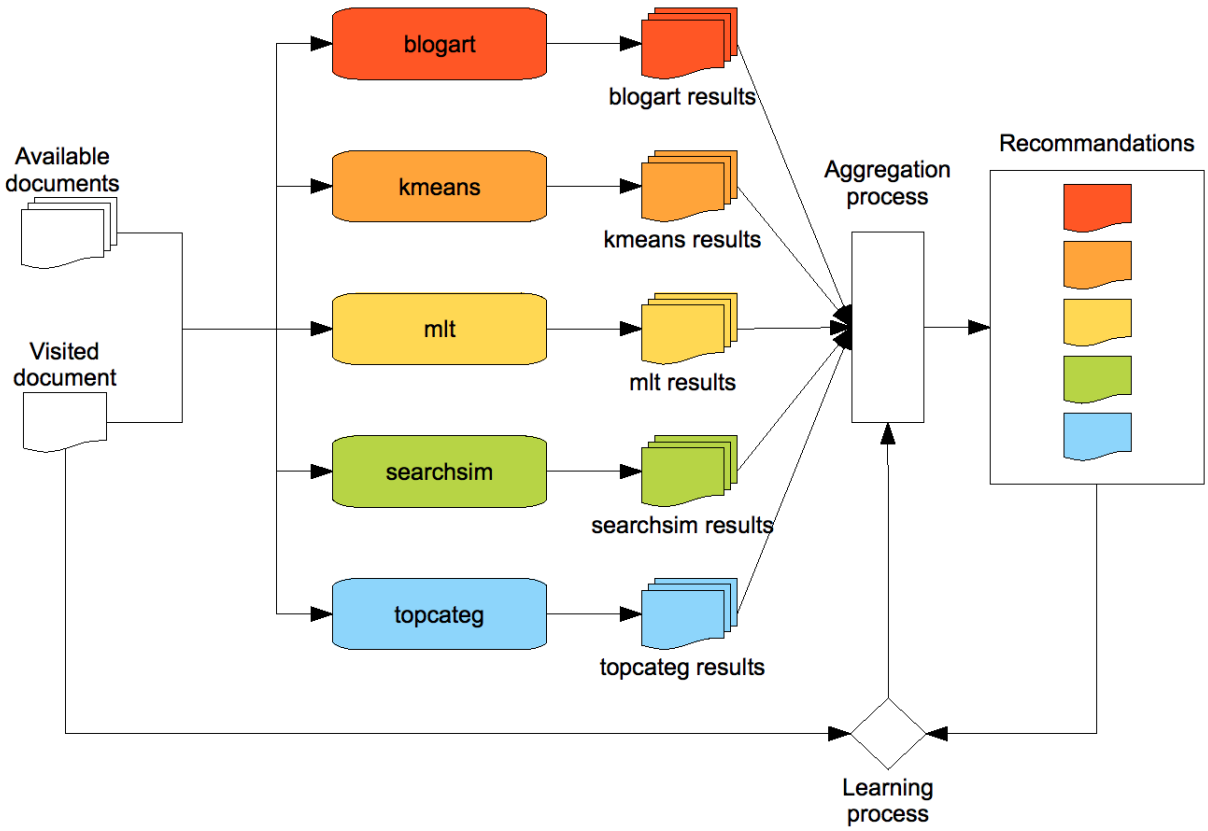

Figure 7. Integration of the learning process in the OverBlog aggregation prototype architecture

(*searchsim*), topical diversity (*kmeans*, *mlt*) or serendipity (*blogart*, *topcateg*) helps to diversify the recommendations and improves the users' satisfaction. We effectively observed a better users' perception of diversity with our RS, without a loss of precision. Indeed, the recommendations resulting from the aggregate similarity measure offer a good balance between accuracy and diversity.

Additionnaly, we promote a framework in which different similarity measures can be combined. One of the main points of this framework is that it is adaptable: any other measure can be added to the framework. It seems that it is worth using approaches that offer serendipity; to this extend, *blogart* seems to be an interesting one. On the other hand, *topcateg* is to be improved.

Our model is not the only one that promote fusing recommandations. Other RS fusion approaches have been proposed in the literature. For example, Schafer *et al.* [58] and Jahrer *et al.* [38] present a "Meta RS". However, when they choose to focus on results shared by the different RS, we instead propose to select the best recommendations from each similarity measure to ensure diversity. We assume that it is important to give a chance to enlarge facets contained in retrieved documents.

When existing approaches focus on designing methods to force diversity in their results (using clustering or MMR), we choose to consider multiple similarity measures to build the recommendation list and ensure diversity. Moreover, it is important that every document may give rise to a wide range of interests for users (a good perception of diversity while keeping a good accuracy level in the recommendation list).

We will direct our future work towards completing the RS architecture to better fit with users' expectations. That is why we will study the learning mechanism to find the proportion of documents coming from every similarity measure, for a given browsed document. As shown on Figure 7, the system may learn the main interests that are important for end-users. To do this, the idea is to use an automatic learning process based on users' feedbacks. We could for example simply initialize the system with equal distribution for each RS (each system contributes equaly to the final list of recommendations), and then increase the proportion of recommendations comming from systems that recommend documents that are more often clicked by the users, and decrease the proportion of recommandation from RS less often considered. Considering the results of the experiments

presented in this paper, we could expect a 80% proportion for topicality systems, and 20% for more original systems. Our future work will also analyze if results are consistent on a real scale experiment using the online blog platform OverBlog when using learning.

REFERENCES

[1] L. Candillier, M. Chevalier, D. Dudognon, and J. Mothe, "Diversity in recommender systems: bridging the gap between users and systems," in *CENTRIC 2011, The Fourth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services*, 2011, pp. 48–53.

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

[3] T. Malone, K. Grant, F. Turbak, S. Brobst, and M. Cohen, "Intelligent information-sharing systems," *Communications of the ACM*, vol. 30, no. 5, pp. 390–402, 1987.

[4] M. Montaner, B. López, and J. De La Rosa, "A taxonomy of recommender agents on the internet," *Artificial intelligence review*, vol. 19, no. 4, pp. 285–330, 2003.

[5] A. Kumar and D. Thambidurai, "Collaborative web recommendation systems-a survey approach," *Global Journal of Computer Science and Technology*, vol. 9, no. 5, 2010.

[6] M. Chevalier, T. Dkaki, D. Dudognon, and J. Mothe, "Recommender system based on random walks and text retrieval approaches," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - Discovery Challenge Workshop (ECML/PKDD - DCW)*, T. Smuc, N. Antulov-Fantulin, and M. Morzy, Eds. http://www.irb.hr/: Rudjer Boskovic Institute, 2011, pp. 95–102.

[7] G. Salton and M. McGill, *Introduction to modern information retrieval*. McGraw-Hill, Inc., 1983.

[8] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 659–666.

[9] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 335–336.

[10] X. Yin, X. Huang, and Z. Li, "Promoting ranking diversity for biomedical information retrieval using wikipedia," *Advances in Information Retrieval*, pp. 495–507, 2010.

[11] A. Chifu and R. T. Ionescu, "Word sense disambiguation to improve precision for ambiguous queries," *Central European Journal of Computer Science*, 2012.

[12] K. Bradley and B. Smyth, "Improving recommendation diversity," in *12th National Conference in Artificial Intelligence and Cognitive Science (AICS-01)*. Citeseer, 2001, pp. 75–84.

[13] G. Adomavicius and Y. Kwon, "Improving aggregate recommendation diversity using ranking-based techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 896–911, 2012.

[14] P. Borlund, "The concept of relevance in ir," *Journal of the American Society for information Science and Technology*, vol. 54, no. 10, pp. 913–925, 2003.

[15] Text REtrieval Conference (TREC) homepage. 04.12.2012. [Online]. Available: http://trec.nist.gov

[16] J. Mothe and G. Sahut, "Is a relevant piece of information a valid one? teaching critical evaluation of online information," *Teaching and Learning in Information Retrieval*, pp. 153–168, 2011.

[17] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *2nd ACM International Conference on Web Search and Data Mining*. ACM, 2009, pp. 5–14.

[18] R. Santos, C. Macdonald, and I. Ounis, "Selectively diversifying web search results," in *19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1179–1188.

[19] Y. Xu and Z. Chen, "Relevance judgment: What do information users consider beyond topicality?" *Journal of the American Society for Information Science and Technology*, vol. 57, no. 7, pp. 961–973, 2006.

[20] F. Radlinski, P. Bennett, B. Carterette, and T. Joachims, "Redundancy, diversity and interdependent document relevance," in *ACM SIGIR Forum*, vol. 43, no. 2. ACM, 2009, pp. 46–52.

[21] C. Ziegler, S. McNee, J. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *14th international conference on World Wide Web*. ACM, 2005, pp. 22–32.

[22] J. He, K. Balog, K. Hofmann, E. Meij, M. Rijke, M. Tsagkias, and W. Weerkamp, "Heuristic ranking and diversification of web documents," DTIC Document, Tech. Rep., 2009.

[23] W. Bi, X. Yu, Y. Liu, F. Guan, Z. Peng, H. Xu, and X. Cheng, "Ictnet at web track 2009 diversity task," DTIC Document, Tech. Rep., 2009.

[24] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *5th Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 281-297. California, USA, 1967, p. 14.

[25] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[26] R. Kaptein, M. Koolen, and J. Kamps, "Result diversity and entity ranking experiments: Anchors, links, text and wikipedia," DTIC Document, Tech. Rep., 2009.

[27] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

[28] N. Lathia, S. Hailes, L. Capra, and X. Amatriain, "Temporal diversity in recommender systems," in *33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR10)*, 2010, pp. 210–217.

[29] G. Cabanac, M. Chevalier, C. Chrisment, and C. Julien, "An original usage-based metrics for building a unified view of corporate documents," in *Database and Expert Systems Applications*. Springer, 2007, pp. 202–212.

[30] D. Dudognon, G. Hubert, J. Marco, J. Mothe, B. Ralalason, J. Thomas, A. Reymonet, H. Maurel, M. Mbarki, P. Laublet, and V. Roux, "Dynamic ontology for information retrieval," in *Adaptivity, Personalization and Fusion of Heterogeneous Information*. CID, 2010, pp. 213–215.

[31] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.

[32] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[33] I. Esslimani, A. Brun, and A. Boyer, "A collaborative filtering approach combining clustering and navigational based correlations," *Web Information Systems and Technologies*, pp. 364–369, 2009.

[34] L. Jabeur, L. Tamine, and M. Boughanem, "A social model for literature access: Towards a weighted social network of authors," in *Adaptivity, Personalization and Fusion of Heterogeneous Information*. CID, 2010, pp. 32–39.

[35] J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, and S. Karouach, "Combining mining and visualization tools to discover the geographic structure of a domain," *Computers, environment and urban systems*, vol. 30, no. 4, pp. 460–484, 2006.

[36] K. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[37] The Amazon website. 04.12.2012. [Online]. Available: http://www.amazon.com

[38] M. Jahrer, A. Töscher, and R. Legenstein, "Combining predictions for accurate recommender systems," in *16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 693–702.

[39] C. Vogt and G. Cottrell, "Fusion via a linear combination of scores," *Information Retrieval*, vol. 1, no. 3, pp. 151–173, 1999.

[40] M. Chevalier, A. Dattolo, G. Hubert, and E. Pitassi, "Information retrieval and folksonomies together for recommender systems," *E-Commerce and Web Technologies*, pp. 172–183, 2011.

[41] C. Clarke, N. Craswell, and I. Soboroff, "Overview of the trec 2009 web track," DTIC Document, Tech. Rep., 2009.

[42] C. Clarke, N. Craswell, I. Soboroff, and A. Ashkan, "A comparative analysis of cascade measures for novelty and diversity," in *4th ACM international conference on Web search and data mining*. ACM, 2011, pp. 75–84.

[43] C. Hayes, P. Massa, P. Avesani, and P. Cunningham, "An on-line evaluation framework for recommender systems," in *Workshop on Personalization and Recommendation in E-Commerce*. Springer Verlag, 2002.

[44] J. Lee, "Analyses of multiple evidence combination," in *ACM SIGIR Forum*, vol. 31. ACM, 1997, pp. 267–276.

[45] The Lemur toolkit for language modeling and information retrieval. 04.12.2012. [Online]. Available: http://www.lemurproject.org/indri

[46] P. Chandar, A. Kailasam, D. Muppaneni, L. Thota, and B. Carterette, "Ad hoc and diversity retrieval at the university of delaware," in *Text REtrieval Conf*, 2009.

[47] W. Zheng and H. Fang, "Axiomatic approaches to information retrieval-university of delaware at trec 2009 million query and web tracks," DTIC Document, Tech. Rep., 2009.

[48] The Apache Hadoop project homepage. 04.12.2012. [Online]. Available: http://hadoop.apache.org

[49] J. Lin, D. Metzler, T. Elsayed, and L. Wang, "Of ivory and smurfs: Loxodontan mapreduce experiments for web search," DTIC Document, Tech. Rep., 2009.

[50] Terrier IR platform homepage. 04.12.2012. [Online]. Available: http://www.terrier.org

[51] The free encyclopedia Wikipedia homepage. 04.12.2012. [Online]. Available: http://www.wikipedia.org

[52] R. McCreadie, C. Macdonald, I. Ounis, J. Peng, and R. Santos, "University of glasgow at trec 2009: Experiments with terrier," DTIC Document, Tech. Rep., 2009.

[53] A. Spink and B. Jansen, "A study of web search trends," *Webology*, vol. 1, no. 2, p. 4, 2004.

[54] P. Pu, L. Chen, and R. Hu, "Evaluating recommender systems from the user's perspective: survey of the state of the art," *User Modeling and User-Adapted Interaction*, pp. 1–39, 2012.

[55] E. Fox and J. Shaw, "Combination of multiple searches," *NIST Special Publication*, pp. 243–243, 1994.

[56] The OverBlog website. 04.12.2012. [Online]. Available: http://www.over-blog.com

[57] Solr open source enterprise search platform. 04.12.2012. [Online]. Available: http://lucene.apache.org/solr

[58] J. Schafer, J. Konstan, and J. Riedl, "Meta-recommendation systems: user-controlled integration of diverse recommendations," in *11th international conference on Information and knowledge management*. ACM, 2002, pp. 43–51.