Intractable problems in novelty and diversity

Simone Santini and Pablo Castells*

Escuela Politécnica Superior, Universidad Autónoma de Madrid

Abstract. Information retrieval's basic problem is retrieving a set of documents *relevant* for a given query. Here, we present three classes of methods that appeared in the liteature, as well as a new one, which is an improvement of the one the three, to retrieve result sets that, in addition to relevance, try to maximize *diversity* and *novelty*. We analyze the complexity of these problems and show that whenever relevance, diversity, and novelty are considered together, the methods are all NP-complete.

1 Introduction

In information retrieval, a user expresses certain information needs through a query, which normally consists of a collection of keywords, and wants to retrieve, from a generally large collection of documents, a sub-set *relevant* to the query. Traditionally, this has been done by evaluating, through opportune algorithms, the *estimated relevance* of each document for the query expressed by the user. The optimal result of size n will then consist of a list of the n document with the highest relevance value, presented in decreasing order of relevance [9]. As neutral and objective as it might seem, this model is based on some fairly strong assumptions regarding relevance and its estimates [10]. Namely the model assumes that relevance is:

i) *topical*—relevant documents are about the same topic as the query;
ii) *independent*—the relevance of a document does not depend on the relevance of the other documents in the collection;
iii) *stable*—relevance does not change over time;
iv) *consistent*—relevance judgments do not depend on who expressed them, that is, different people will agree on which documents are relevant for which query;
v) *complete*—all the documents have a relevance judgment.

Starting towards the end of the 1990s, these assumptions have been questioned, and different models of information retrieval based on the rejection of some of them have been proposed (we call these: *non-Robertsonian* information retrieval models). In particular, *topicality* has been amended with the introduction of sub-topics that are judged independently, thereby introducing a structure into the simple notion of topicality [14], the completeness assumption has been relaxed by inferring the relevance of missing

* This work was supported in part by the *Ministerio de Educación y Ciencia* under the grant N. MEC TIN2008-06566-C04-02, *Information Retrieval on different media based on multidimensional models: relevance, novelty, personalization and context.*

documents in various ways, as done by [3], or by carefully selecting for which documents a judgment should be obtained, as in [5].

In this paper, we are interested in non-Robertsonian information retrieval systems and evaluation metrics that dispense of the *independence* assumptions, specifically in metrics that assume that the relevance of a document depends on that of the documents that are shown together with it.

The general idea is the following. Suppose we are entering a fairly generic query, for instance, we simply introduce the word *manhattan*. From the query alone, it is not quite clear what we refer to: the island in New York, the cocktail with the same name, the film by Woody Allen, or the native tribe from which the Dutch claimed to have bought the island. The query is *ambiguous* in that it has several mutually exclusive *interpretations*. The interpretations are mutually exclusive because, in general, a user interested in information about New York, will not be interested in information about the indian tribe or the cocktail, and vice versa. Let us now fix our attention on an interpretation, for example in the part of New York. It is still not completely clear what we are after: Manhattan is a pretty broad subject, and there are many different *aspects* of it in which we might be interested. We might be interested in the history of Manhattan, or in its urban planning. We might be interested in its buildings, or in how to get around it by bus. The query is, in other words, *underspecified*, that is, it has a number of possible *aspects* of interest. Unlike interpretations, aspects are not necessarily mutually exclusive: documents about urban planning may be deemed interesting also by a person interested mostly in the history of Manhattan.

To cope with ambiguity, the results of a query should be *diverse*, that is, they should cover more that one interpretation of the query, so as to give all users some useful results. In order to cope with underspecification, each results in the list should be, at least in part, *novel*, that is, it should give the user some information that the previous documents did not give.

Despite being often cited together, diversity and novelty are quite different concepts, as one of them deals with ambiguity (the same words can refer to different, unrelated areas), the second with underspecification (a query, by its very nature, never specifies completely and exactly what document does the user want—if it did, we wouldn't be doing information retrieval but data bases).

In the last few years, several methods have been proposed to maximize the diversity and novelty of result sets in information retrieval, based on formal definitions of these two concepts. Despite the great interest of these methods, the complexty of the resulting optimization problem has rarely been studied. To the best of our knowledge, the only available result is the NP-completeness of the DIVERSIFY(n) problem defined in [1].

Our purpose in this paper is twofold. First, we present a general overview of retrieval techniques that optimize diversity and/or novelty in information retrieval. In so doing, we shall also propose a new technique that improves on those presented in the literature. Second, we show that all these problems are NP-complete, that is, intractable. The result might at first look disappointing but, as we shall argue in the conclusions, it opens a plethora of interesting problems for researchers to study.

2 The Portfolio theory

The *portfolio theory of information retrieval*, introduced by [11] is a very recent approach to diversity, based on ideas from the portfolio theory of investment of [6].

Suppose we have a set of documents returned by a query, $D = \{d_1, \ldots, d_n\}$ with a *relevance* score r_i associated to each document. The result of a query is an ordered list of documents. We might want to associate a weight to each position of the list, to model the fact that the first positions are more desirable than the last ones. So we define a set of descreasing normalized weights $w_1 > w_2 > \cdots > w_n$ with $\sum_{i=1}^{n} w_i = 1$. We can define the *relevance score* of the list as

$$R_n = \sum_{i=1}^{n} w_i r_i \tag{1}$$

It is easy to show that the weight ordering entails that R_n is maximum when $r_1 > r_2 > \cdots > r_n$. Relevance, however, is subject to uncertainty, since different users might grade the same document differently (due to the ambiguity of the query). We can then model relevance scores as stochastic variables.

Define $\mathbb{E}[r_i]$ as the mean of r_i, and $\Sigma[R_n]$ as the covariance matrix of the ranking R_n, where $\sigma_{i,i}$ is the variance of r_i and $\sigma_{i,j}$ the covariance of r_i and r_j. With these definitions, we can introduce the expected value and the variance of the relevance score of the list as:

$$\mathbb{E}[R_n] = \sum_{i=1}^{n} w_i \mathbb{E}[r_i] \tag{2}$$

$$\sigma[R_n] = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{i,j} \tag{3}$$

We do, of course, want a result set with a high value of $\mathbb{E}[R_n]$, since we want something that, on average, will be relevant. However, if the variance $\sigma[R_n]$ is high, we are assuming a certain amount of risk: on average, the ranking R_n will give us a good result, but the high variance tells us that the score of the result will be very spread so, if on average the results are good, there is a significant fraction of users that will consider them poor.

So, while we maximize the mean relevance, we'd better keep an eye on the variance of the results, and try to keep it small. One standard way of doing this is by minimizing a linear combination of mean and variance:

$$O_n = \mathbb{E}[R_n] - b\sigma[R_n] \tag{4}$$

Where b is positive a weight parameter. Mathemathically, the minimization problem (4) is equivalent to the following [2]:

> PORTFOLIO(K): given a set D of n documents and the random variables r_i, find a subset $S \subseteq D$ with $|S| = k$ that solves
>
> $$\min\ \sigma[R_n] \ \text{ subject to } \ \mathbb{E}[R_n] = m \tag{5}$$

where

$$R_n = \sum_S w_i r_i \tag{6}$$

With regard to this, we can prove the following:

Theorem 1. *The optimization problem* PORTFOLIO(N) *is NP-hard.*

Proof. We prove that the corresponding decision problem is NP-complete: given a set of relevance values $R = \{r_1, \ldots, r_n\}$ and values m, q, determine whethere there exists a subset $R_n \subseteq R$, of size n such that $\sigma[R_n] \leq q$ and $\mathbb{E}[R_n] = m$.

We will prove that the decision problem is NP-complete by reduction from a modified form of MAXIMUM 2SAT. We are given a set $U = \{u_1, \ldots, u_n\}$. The *literal* l_i is either the variable u_i or its negation $\bar{u}_i$. We are also given a set of disjunctive clauses $C = \{c_1, \ldots, c_m\}$. Each clause is of the form $c_k = l_{k_1} \vee l_{k_2}$. Given a number $K' \leq m$ the problem is to determine whether there is a truth assignment to the variables of U that satisfies at least K clauses. The problem is NP-complete unless $K' = m$. We consider an alternative form of the same problem. Using De Morgan's theorem, we can write $c_k = \neg(\bar{l}_{k_1} \wedge \bar{l}_{k_2})$. Defining the set of clauses $\tilde{C} = \{\tilde{c}_1, \ldots, \tilde{c}_m\}$ with $\tilde{c}_k = \bar{l}_{k_1} \wedge \bar{l}_{k_2}$, the problem is equivalent, given a number K ($K = m - K'$) to determine a truth asssignment to the variables of U such that *at most* K clauses in $\tilde{C}$ are true.

Given a problem in this modified form, we define a set of documents $D = \{d_1, \ldots, d_m, d_{m+1}, \ldots, d_{2m}\}$. The documents $d_1, \ldots, d_m$ correspond to the positive literals of the m variables, the documents $d_{m+1}, \ldots, d_{2m}$ to the negated literals. We give a relevance $r_i = 1$ to each document. We then set the correlation coefficients $c_{ij} = n$ if the conjunction of literals l_i and l_j makes n formulas true[1], and $c_{i,i+m} = M > 2(n+m)$. All the weights w_i are set to 1. Now we solve the optimization problem

$$\min\ \sigma[R_n] \ \text{ subject to } \ \mathbb{E}[R_n] = n \tag{7}$$

If $\min\ \sigma[R_n] < K$ then the decision problem has a solution. Note that the condition $c_{i,i+m} = M > 2(n+m)$ prevents us to find a solution in which the same variable is given two different truth values so that, for instance, if the literal l_i is part of the solution, the literal l_{i+m} can't be.

A model more or less along the same lines was used by [8] for diversifying the results of web searches. One difference, not a terribly relevant one, is that [8] use the form (5) of the optimization problem. A much more interesting difference is that for Rafiei *et al.* the set of documents is fixed, and the unknown in the optimization problem (5) are the (real) values of the weights w_i, a difference that simplifies the problem transforming it from a discrete optimization to a continuous one that can be solved using standard numerical algorithms.

[1] In this formulation, we assume that the coefficients can take arbitrary values, while in reality they are constrained to be in the interval $[-1, 1]$. It is easy to normalize them in such a way that they will satisfy the constraint, but we will not consider the normalization to work with more intuitive values.

3 Diversity with categories

In [1], a more structured problem is considered, in which documents are assumed to belong to one or more *categories* (a better name for these entities would be *topics*, but here we shall retain the nomenclature of the original paper), and that queries are also about categories.

Let $C(q)$ be the set of categories to which query q belongs, and $C(d)$ the categories of document d. The two may or may not overlap. The *user intent* upon issuing query q is represented as a probability distribution over the categories, conditioned by the text of the query. That is, $P(c|q)$ is the probability that, having issued the query q, the user will be interested in documents of category c. We assume complete knowledge, that is, $\sum_{c \in C(q)} P(c|q) = 1$.

The relevance of a document d, is a function not only of the query, but of the category as well. If a query is "about" several categories, the same document d can be relevant for some of them and irrelevant for some others. Relevance is modeled as a probability, and $V(d|c,q)$ is the *likelihood* that document d be relevant for category c given the query q, that is, given the user intent (q,c). An independence assumption is made: the likelihood that two documents satisfy the same user intent is simply the product of their individual likelihoods.

The basic idea of the paper is to develop a sorting criterion that tries to satisfy categories depending on their probability of being the category of interest for a given query but that also "discounts away" a category if it has been adequately satisfied. If a very relevant document about a given category has already been retrieved, there is not much to be gained from retrieving more documents about the same category.

The *gain* that we obtain by adding a new document to the list of results decreases when the categories covered by that documents have already been covered by other documents in the set. The retrieval process should maximize the probability that the average user will find at least one interesting document in the result set. Suppose we have a query q, and we fix a category c related to the query. We are given a set S of documents; the independence assumption entails that the probability that *no* document in S be relevant for category c (given query q) is $\prod_{d \in S}(1 - V(d|c,q))$ and the probability that at least one document be relevant is $1 - \prod_{d \in S}(1 - V(d|c,q))$. In order to satisfy the average user, we weight these values with the probability that category c be relevant given query q, and sum over all categories (we can do this because of our hypothesis of complete knowledge). So, we have to determine the set S that maximizes

$$P(S|q) = \sum_{c} P(c|q) \left[1 - \prod_{d \in S} (1 - V(d|c,q)) \right] \tag{8}$$

If, for a given category, there is a document $\hat{d}$ for which $V(\hat{d}|c,q)$ is very high, adding more documents of the same category will not increase the objective function by much, because every increase will be multiplied by the factor $1 - V(\hat{d}|c,q)$, which is small. In the extreme case of $V(\hat{d}|c,q) = 1$, we have $1 - \prod_{d \in S}(1 - V(\hat{d}|c,q)) = 1$ regardless of the characteristics of the other documents, so, once we have retrieved the *perfect* document for a category, there is no advantage in retrieving more documents for the same category: we can move on to other topics.

Categories are weighted through their probability $P(c|q)$, entailing that it will be more convenient to spend more of our *document budget* (the size of S) to serve well a common category, even if this comes at the expense of a less common category. In this sense, the quantity $P(S|q)$ is the probability that the average user entering the query q will find at least a relevant document in the set S. The problem that the authors call DIVERSIFY(K) requires, given a query q, finding the set S with $|S| = k$ that maximizes $P(S|q)$. Unfortunately for us [1] proves the following result:

Theorem 2. DIVERSIFY(K) *is NP-complete*

The proof is based on a reduction from MAX COVERAGE, and the interested reader can find it in [1].

Note that DIVERSIFY(K) does not assume aby ordering of the documents, since S is a set. This is a consequence of the lack of a user model in the problem: we assume that, upon receiving the set of k documents, a user is equally likely to look at any one of them.

* * *

It is not too hard to improve the method by using one of the standard user models. The simplest one considers a user that, after having looked at a document in position k that doesn't satisfy her needs, gives up the search with probability $1 - \beta$, and sets to analyze the next document in a list with probability β. We'll call this the *geometric* user model. In this model, the probability that the user will not be bored before reaching the kth document (independently of the interest in the documents) is β^{k-1}. Given an (ordered) *list* of documents $S = [d_1, \ldots, d_n]$, we are then interested in the probability that the first interesting document for category c be found in the kth position of the list. This probability is given by

$$V(d_k|c,q) \prod_{j=1}^{k} (1 - V(d_j|c,q)) \tag{9}$$

So, a user interested in category c, will find this document with probability

$$\sum_{k=1}^{n} \beta^k V(d_k|c,q) \prod_{j=1}^{k} (1 - V(d_j|c,q)) \tag{10}$$

and the average user will find an interesting document with probability

$$P'(S|q) = \sum_{c} P(c|q) \sum_{k=1}^{n} \beta^k V(d_k|c,q) \prod_{j=1}^{k} (1 - V(d_j|c,q)) \tag{11}$$

The problem RANKED-DIVERSIFY(K) is then defined as follows:

> Given a set of n documents D, a query q and a set of categories C, determine the list S of elements of D, with $|S| = k$ that maximizes the value $P'(S|q)$ as given in (11).

Given the way we have derived the problem, it seems obvious that there should be a relation between DIVERSIFY(K) and RANKED-DIVERSIFY(K), and that RANKED-DIVERSIFY(K) should be something of a harder version of DIVERSIFY(K). This is indeed the case, as we shall see shortly. Before doing so, however, we need a technical lemma.

Lemma 1. *Let $[v_1, \ldots, v_n]$ a list of values, $v_i \in \mathbb{R}$. Then for all $k \leq n$ it is*

$$\sum_{u=1}^{k} v_u \prod_{j=1}^{u-1}(1 - v_j) = 1 - \prod_{u=1}^{k}(1 - v_u) \tag{12}$$

Proof. We prove the lemma by induction on k. For $k = 1$ the lemma reduces to $v_1 = 1 - (1 - v_1)$, which is obvious.

Suppose now the lemma is true for $k - 1$ and write (12) as

$$\sum_{u=1}^{k-1} v_u \prod_{j=1}^{u-1}(1 - v_j) + v_k \prod_{j=1}^{k-1}(1 - v_j) = 1 - \prod_{u=1}^{k-1}(1 - v_u)(1 - v_k) \tag{13}$$

Set

$$A = \sum_{u=1}^{k-1} v_u \prod_{j=1}^{u-1}(1 - v_j) = 1 - \prod_{j=1}^{k-1}(1 - v_j) \tag{14}$$

(the two are equal because of the inductive hypothesis), so that (12) becomes

$$A + v_k(1 - A) = 1 - (1 - A)(1 - v_k) \tag{15}$$

which is trivially true.

We show the relation between the two problems while proving the following theorem:

Theorem 3. RANKED-DIVERSIFY(K) *is NP-complete.*

Proof. We prove the theorem by reducing DIVERSIFY(K) to RANKED-DIVERSIFY(K).

Let an instance of DIVERSIFY(K) be given with the values $P(c|q)$ and $V(d|c, q)$, the size of the data base n, and the target set size k. We build a corresponding instance of RANKED-DIVERSIFY(K) by setting $\beta = 1$. In this case, the objective function of RANKED-DIVERSIFY(K) becomes

$$P'(S|q) = \sum_{c} P(c|q) \sum_{u=1}^{k} V(d_u|c, q) \prod_{j=1}^{u}(1 - V(d_j|c, q)) \tag{16}$$

because of lemma 1, we can rewrite this function as

$$\sum_{c} P(c|q)(1 - \prod_{j=1}^{k}(1 - V(d_j|c, q))) \tag{17}$$

which is the objective function of DIVERSIFY(K), so solving RANKED-DIVERSIFY(K) will solve DIVERSIFY(K).

* * *

It seems intuitively plausible that there should be some connection between DIVERSIFY(K) and PORTFOLIO(K) since both start with the same idea: minimize the probability that the average user will find nothing interesting in the result set. Despite this common idea, the two methods are based on different assumptions, which are reflected in the difference between the functions that are being maximized in the two cases. In the case of PORTFOLIO(K) there is no concept of categories, so we should take a special case of (8) in whcih there is only one category. In this case (8) becomes $P'(S|q) = 1 - \prod_{d \in S}(1 - V(d|q))$, and the problem DIVERSIFY(K) is solved by maximizing P' or, equivalently, by minimizing

$$P''(S|q) = \prod_{d \in S}(1 - V(d|q)) \tag{18}$$

We must note, however, that with this change the character of the two problems has changed quite drastically. While PORTFOLIO(K) is still concerned with diversity (trying to reduce the probability that the average user be dissatisfied), DIVERSIFY(K) is now concerned with the full exploration of a single topic. The value $V(d_i|q)$ is the likelihood that document d_i be considered relevant for q. In the portfolio theory, this likelihood corresponds to the average $\mathbb{E}[r_i]$, so we can rewrite P'' as

$$P''(S|q) = \prod_{d_i \in S}(1 - \mathbb{E}[r_i]) \tag{19}$$

We can approximate this product using the equality:

$$\prod_{i=1}^{n}(1-x_i) = 1-\sum_{i=1}^{n} x_i+\sum_{j>i} x_i x_j+o(x_i^3) = 1-\sum_{i=1}^{n} x_i+\frac{1}{2}\sum_{i,j} x_i x_j-\frac{1}{2}\sum_{i=1}^{n} x_i^2+o(x_i^3) \tag{20}$$

So, we can formulate DIVERSIFY as the minimization of (20) or, equivalently, as

$$\max O_A = \max\left[\sum_{i=1}^{n}\mathbb{E}[r_i] - \frac{1}{2}\sum_{i,j=1}^{n}\mathbb{E}[r_i]\mathbb{E}[r_i] + \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}^2[r_i]\right] \tag{21}$$

On the other hand, PORTFOLIO(K) minimizes (4), that is, setting $w_i = 1$, it solves

$$\max O_B = \max\left[\sum_{i=1}^{n}\mathbb{E}[r_i] - b\sum_{i,j=1}^{n}\mathbb{E}[r_i r_j] + b\sum_{i,j=1}^{n}\mathbb{E}[r_i]\mathbb{E}[r_j]\right] \tag{22}$$

The reason to set $w_i = 1$ is that DIVERSIFY(K) works on sets, not on ordered lists, so there is no reason to distinguish between positions in the output, and all the weights can be given the same value. We want to analyze the difference $O_B - O_A$. The two values are most similar when $b = -1/2$, in which case

$$O_B - O_A = -\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}^2[r_i] + \frac{1}{2}\sum_{i,j=1}^{n}\mathbb{E}[r_i r_j] = \frac{1}{2}\left[\sum_{i=1}^{n}\sigma_i^2 + \sum_{i\neq j}\mathbb{E}[r_i r_j]\right] \tag{23}$$

Two interesting considerations can be drawn from this comparison. The first is that, since $b < 0$, DIVERSIFY(K) adopt a *risk-loving* strategy [11]. This is consistent with the different assumptions of the two methods. In DIVERSIFY(K), at least in this case, we are considering a single category, that is, all documents are considered *on topic*, and the interest of the method is to maximize the "spread" of the documents, to explore the different aspects of this topic. Second, PORTFOLIO(K) seems, *coeteris paribus* to favor documents with high variance (high variances make the term $O_B - O_A$ larger) over documents with high correlation. This is also consistent with the assumptions of the two methods: DIVERSIFY(K) assumes high correlations (after all, all documents are about the same topic) as long as every document carries some new information, while in the case of PORTFOLIO(K) low correlation is preferred to cover a larger number of topics and so minimize risk.

4 Considering interaction

The systems considered so far have worked with *one shot* queries so that they had to consider diversity and novelty together, blurring somehow the difference between the two and resorting to statistical considerations to satisfy the "average" user. This is a natural way to pose the problem from the point of view of the server, which must balance the answer considering the different needs of different users.

Interaction offers a way to take the single person's perspective into account. Xu and Yin, in [13], systemathize the rôle of interaction in the light of the recent developments in novelty and diversity.

They operate a quadripartite division of possible systems along two axes. The first axis is presentation, and systems are divided as having *compensatory* or *step* presentation. In a compensatory presentation, topicality and novelty are considered together in order to provide a composite relevance score, and the result list is then created based on this score. In a *step* system, topicality is considered first as a gauge: only documents that score above a certain edge of topicality are retained. Novelty is considered next, and is used to reorder (and, possibly, to filter again) the set of documents that have passed the gauge of topicality.

The second axis deals with interaction, and distinguishes between *undirected* and *directed* systems. Undirected systems are those that we have called "one shot:" they receive a query and return a list of results, returning at each position documents that minimize redundancy with those already returned. Directed systems receive an input from the user indicating in which areas she wants the search to continue.

Users have several criteria in mind when they talk about quality of results, among which the most prominent are *topicality*, *novelty*, *ease of understanding*, *reliability*, and *scope*, although topicality seems to be the most relevant [7]. These findings form the foundations of the methodological division operated by [13]. For example, if a document is off-topic, all other factors are irrelevant for judgment [12]. This property justifies the study of *step* systems, in which documents are ranked only if they are beyond a certain threshold of topicality. On the other hand, computing practicioners don't like arbitraty thresholds, especially when the sensitivity of the system with respect to their value is not easily evaluated, a circumstance that makes it sensible to evaluate compen-

satory system as a more practical and robust solution. As we shall see, this practicality comes with a price: undirected system use less information about the user and the problems they involve are computationally harder.

4.1 User profiles and similarity

A user is characterized by two profiles: the *topicality profile* and the *novelty profile*. Both profiles are dyamic, and are updated as part of a person's interaction with the system. The overall document model is that of a vector space: a document d is a vector in a suitable Euclidean space W whose axes represent words or combinations of words. Each query, indexed by the query order t, is a *round* during which the results are returned and analyzed (by the user), and the profiles are updated. Let D_t be the set of documents examined at time t. Suppose that, to each $d \in D_t$, the user has assigned a *topicality score* $T[d] \in [0,1]$ If $P_{t-1}^T \in W$ was the topicality profile of the user before round t, then the topicality profile after round t is

$$P_t^T = P_{t-1}^T + \frac{1}{|D_t|} \sum_{d \in D_t} d \cdot T[d]. \tag{24}$$

Note that, formally, the topicality profile is a (virtual) document that contains all the topics of interest for the user.

Directed (viz. interactive) systems add to this a *novelty profile*. Here we assume that at iteration t the user will mark some documents as "novel", and these documents are combined to form a term vector P_t^N, which constitutes the novelty profile, at time t, for the user.

Suppose that the user has marked a sample set of documents Q_t as either novel or not novel. We want to use these documents to build an *instantaneous* novelty profile. Our problem is how to go from the set of judged documents Q_t to the term vector $P_{Q_t}^N$. For this, we need a weighting scheme that assigns a weight to each word in Q_t. Xu and Yin reject for this purpose the use of the common TF-IDF, as they consider it better at differentiating topics than at differentiating between documents on the same topic. Rather, they use the probabilittic measure F4 of [9]. The reader is referred to [13] for details.

The novelty profile that is actually used in the systems that we shall consider is a *smoothed* version pf $P_{Q_t}^N$ defined as

$$P_t^N = (1-\nu)P_{t-1}^N + \nu P_{Q_t}^N \tag{25}$$

where $\nu \in [0,1]$ is the smoothing parameter that determines the dynamics of the profile.

4.2 Undirected-compensatory systems

The first class of systems that we consider is that of *undirected, compensatory* systems. In these systems, topicality and novelty are combined in a single measure that is used to order the documents. Novelty is undirected, so the user gives no indication on which documents are novel and which are not, and the system increases novelty simply by

trying to reduce redundancy. The topicality of a document d, here as in other systems, is estimated simply as the similarity between the topicality profile and d_i, given by a suitable similarity measure $s(d_i, P^T)$. Measuring novelty in an undirected context requires a measure of the redundancy of a document d placed in a set of documents D, which the authors characterize as the maximum similarity between d and any other document in the set:

$$\text{Rd}(d_i|D) = \max_{d' \in D} s(d, d') \tag{26}$$

The relevance of a document is then a weighted sum of its topicality discounted by its redundancy:

$$r(d|D) = \alpha s(d, P^T) - (1 - \alpha)\text{Rd}(d|D) \tag{27}$$

with $0 \leq \alpha \leq 1$. The problem is then to find the set of documents D with the largest values of r. Note that this formulation of the problem is a simple modification of the MMR model of [4] in which the topicality profile P^T replaces the original query. Formally, the problem UC(K) can be defined as:

> UC(K): given a set of documents D with $|D| = n$, and a topicality profile P^T, find a subset $S \subseteq D$, with $|S| = k$ such that $\sum_{d \in S} r(d|S)$ is maximal.

4.3 Undirected-step systems

Adapting the previous system to step-wise relevance judgment is quite easy: one has simply to define a relevance measure that filters the documents by topicality before ordering them by decreasing redundancy. The relevence $r(d|D)$ is then defined as

$$r'(d|D) = \begin{cases} 0 & \text{if } s(d, P^T) \leq s^* \\ 1 - \text{Rd}(d|D) & \text{if } s(d, P^T) > s^* \end{cases} \tag{28}$$

where s^* is a suitable relevance threshold. Setting the threshold s^* might be a problem, and it might result in results sets of widely varying size depending on the query. [13] discuss the possbility of replacing the cut-off based on the degree of topicality with one based on the number of results, for instance consiering always the 20 most topical documents and apply the redundancy measure to them. The resulting problem can be formalized as

> US(K): given a set of documents D with $|D| = n$, and a topicality profile P^T, find a subset $S \subseteq D$, with $|S| = k$ such that $\sum_{d \in S} r'(d|S)$ is maximal.

4.4 Directed-compensatory systems

Directed systems use the novelty profile built by the user during the interaction with the system. We must note that, in spite of its superficial resemblance, the dynamics of the novelty profile (25) is of a different nature than the dynamics of the topicality profile (24). The latter is built historically, through a number of interactions with the system, while the former is created through selections done while answering a single query. The dynamics (25) is therefore a rapid one (it is–or may be–restarted with every query),

while (24) is a slow one, updated as a result of various interactions with the system. In a compensatory system, the interactive relevance of a document d ($i(d)$) after t iterations is simply a weighted sum of its topicality (similarity with the topicality profile of the user) and its novelty (similarity with the novelty profile):

$$i_t(d) = \gamma s(d, P^T) + (1 - \gamma) s(d, P_t^N) \tag{29}$$

The corresponding optimization problem is

> DC(K): given a set of documents D with $|D| = n$, a topicality profile P^T, and a novelty profile P_t^N, find a subset $S \subseteq D$, with $|S| = k$ such that $\sum_{d \in S} i(d)$ is maximal.

4.5 Directed-step problems

The corresponding stepwise problem is obtained as in the case of undirected methods: we first filter by topicality, retaining only the documents whoe topicality is beyond a certain threshold, and then order them by novelty. This entails using a relevance score equal to

$$i'_t(d) = \begin{cases} 0 & \text{if } s(d, P^T) \leq s^* \\ s(d, P_t^N) & \text{if } s(d, P^T) > s^* \end{cases} \tag{30}$$

The corresponding optimization problem is

> DS(K): given a set of documents D with $|D| = n$, a topicality profile P^T, and a novelty profile P_t^N, find a subset $S \subseteq D$, with $|S| = k$ such that $\sum_{d \in S} i'_t(d)$ is maximal.

4.6 Complexity of the problems

Two of the four problems presented so far can be solved efficiently. In the problems DC(K) and DS(K), the relevance of a document is independent of the presence of the other documents, so all these problems can be solved quite easily by sorting the set D by relevance and taking the k most relevant documents. That is, these two problems have complexity $O(n \log k)$.

In the case of UC(K), things are more complicated due to the presence of the term $\max_i s(d, d_i)$, which causes the relevance of a document to depend on the relevance of the other documents in the set. This, it turns out, is enough to make the problem intractable:

Theorem 4. UC(K) *is NP-complete.*

Proof. We prove the theorem with a reduction from EXACT COVER BY 3-SETS. The problem is as follows: given a set X, with $|X| = n = 3q$, and a collection C of subsets $C_k \subseteq X$ with $|C| = 3$, find a sub-collection $C' \subseteq C$ such that each element of X occurs in exactly one member of C'.

Given an instance of EXACT COVER BY 3-SETS, we reduce it to UC as follows. We order arbitrarily the elements of X as $[u_1, \ldots, u_n]$ and, for every $C_k = \{u_{k1}, u_{k2}, u_{k3}\}$

we create a document d_k with the weights in the dimensions $k1$, $k2$, and $k3$ equal to 1 and all the other weights equal to 0. The profile P^T is set to a vector with all 1's, and α is set to $1/n$. We show that EXACT COVER BY 3-SETS is solvable if and only if UC(Q) has a solution with $\sum r(d|D) = 1$.

In order to prove this, we shall write the objective function as

$$\sum_{k=1}^{q} r(d_k|C') = \alpha \sum_{k=1}^{q} s(d_k|P^T) - (1-\alpha) \sum_{k=1}^{q} \max_{d \in C'} s(d_k, d) = A - B \tag{31}$$

where C' is the set of q documents that we are considering. Note that each one of the q documents has exactly three non-zero weights, and that the value of these weights is 1, so

$$A = \alpha \sum_{k=1}^{q} s(d_k|P^T) = \frac{1}{n} \sum_{k=1}^{q} 3 = 1 \tag{32}$$

independently of the set C'.

Suppose now that there is a set C' of q sets $\{C_1, \ldots, C_q\}$ that constitutes a solution of EXACT COVER BY 3-SETS. Let the q associated documents be $\{d_1, \ldots, d_q\}$. For each pair C_k, C_h it is $C_k \cap C_h = \emptyset$, so $s(d_k, d_h) = 0$. Therefore, all terms in B are zero and $\sum_{k=1}^{q} r(d_k|C') = A = 1$.

Conversely, if a solution with $\sum_{k=1}^{q} r(d_k|C') = 1$ exists, it must be $B = 0$, so the documents $d_1, \ldots, d_q$ have no axis in common. The sets $C_1, \ldots, C_q$ corresponding to the documents are therefore disjoint. Moreover, since $A = 1$, for each d_k, it is $s(d_k|P^T) = 3$ and since all the d_k are disjoint, no axis is counted more than once so in order to be $A = 1$ there must be a document d_k with a 1 on each axis, proving that the C_k cover X.

A similar reduction proves the following:

Theorem 5. US(K) *is NP-complete.*

We omit the proof, which is almost identical to that of the previous theorem. Note that the first summation on the left-hand side of (31) is always 1, so every value of s^* less than one will make all the documents pass to the following evaluation and the second term of (31), which is the one we are really optimizing, is essentially equal to the second line of (28).

5 Conclusions

Diversity and novelty are desirable properties of a result set in information retrieval. Diversity is an useful property to deal with ambiguous queries, in which there are several mutually incompatible *interpretations*; novelty is useful for underspecified queries, which present several different *aspects* of potential interest to the user. If we aim at increasing novelty and/or diversity in a result set, then the result of a query is no longer a list of the document with the highest score, where the score is computed independently for each document: the score of a document will depend on which other documents are

in the result est. This dependence makes the problem harder. Here, we have proved that all the major rpproaches to the retrieval of novel and diverse result sets are NP-complete. A new method that we have developed in the paper (RANKED-DIVERSITY(n)) also gives rise to an NP-complete problem.

The diverse nature of these methods hints strongly at the fact that virtually all such methods (at least those that in [13] are called *continuous*) may be intractable. This is an important conclusion, and it entails that researchers should focus on the study of approximate, fast methods rather than trying to solve the optimization problem. The result, after all, does not eliminate the possibility of going very close to the optimum with a polynomial method.

References

1. Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Leong. Diversifying search results. In *Proceedings of WDSM '09*. ACM, 2009.
2. S. Boyd and L. Vanderberghe. *Convex Optimization*. Cambridge, University Press, 2004.
3. S. Büttcher, C. L. Clarke, P. C. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgments. In *Proceedings of the 30th International ACM SIGIR Conference in Research and Developmens in Information Retrieval*, pages 63–70. ACM, 2007.
4. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordeing documents and producing summaries. In *Proceedings of the 21th International ACM SIGIR Conference in Research and Developmens in Information Retrieval*. ACM, 1998.
5. B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th International ACM SIGIR Conference in Research and Developmens in Information Retrieval*, pages 268–75. ACM, 2006.
6. H. Markowitz. Portfolio selection. *Journal of Finance*, 1952.
7. S. Mizzaro. Relevance: the whole history. *Journal of the American Society for Information Science and Technology*, 48(9):810–32, 1997.
8. Davood Rafiei, Krishna Bharat, and Anand Shukla. Diversifying web search results. In *Proceedings of WWW 2010, the International Conference on the World Wide Web*. ACM, 2010.
9. S. E. Robertson and K. Spark-Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–46, 1976.
10. T. Saracevic. Relevance: a review of the literature and a framework for thinking on the notion of information science. *Journal of the American Sociery of Information Science and Technology*, 58(13):2126–44, 2007.
11. J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference in Research and Developmens in Information Retrieval*. ACM, 2009.
12. P. Wang and D. Soergel. A cognitive model of document use during a research project. study I: document selection. *Journal of the American Society for Information Science and Technology*, 49(2):445–63, 1998.
13. Yunjie Xu and Hainan Yin. Novelty and topicality in interactive information retrrieval. *Journal of the American Society for Information Science and Technology*, 59(2):201–15, 2008.
14. C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metric for subtopic retrieval. In *Proceedings of the 26th International ACM SIGIR Conference in Research and Developmens in Information Retrieval*, pages 10–7. ACM, 2003.

Madrid, April 2011