PEOPLE RETRIEVAL LEVERAGING TEXTUAL AND SOCIAL DATA

Amin Mantrach and Jean-Michel Renders
Xerox Research Centre Europe, Chemin de Maupertuis, 38240 Meylan, France

Keywords: Social media mining, Information retrieval, Social retrieval, Data fusion, Data aggregation, Multi-view problems, Multiple graphs, Collaborative recommendation, Similarity measures, Pseudo-relevance feedback.

Abstract: The growing importance of social media and heterogeneous relational data emphasizes to the fundamental problem of combining different sources of evidence (or modes) efficiently. In this work, we are considering the problem of people retrieval where the requested information consists of persons and not of documents. Indeed, the processed queries contain generally both textual keywords and social links while the target collection consists of a set of documents with social metadata. Traditional approaches tackle this problem by early or late fusion where, typically, a person is represented by two sets of features: a word profile and a contact/link profile. Inspired by cross-modal similarity measures initially designed to combine image and text, we propose in this paper new ways of combining social and content aspects for retrieving people from a collection of documents with social metadata. To this aim, we define a set of multimodal similarity measures between socially-labelled documents and queries, that could then be aggregated at the person level to provide a final relevance score for the general people retrieval problem. Then, we examine particular instances of this problem: author retrieval, recipient recommendation and alias detection. For this purpose, experiments have been conducted on the ENRON email collection, showing the benefits of our proposed approach with respect to more standard fusion and aggregation methods.

1 INTRODUCTION

With the growing importance of social media and, more generally, of contents that mix heterogeneous types of relational data, the fundamental problem of merging different sources of evidence (or modes) into a common framework is now becoming crucial. We are considering here the problem of people retrieval, which could be considered as an extension of the classical document retrieval paradigm, but where the requested information consists of persons rather than documents. The general case consists of processing a query containing both textual keywords and social links (in other words: a set of key persons), while the target collection is assumed to contain documents (or other textual entities) with social metadata. These social metadata could consist of author/recipient relationships (like in email collections), co-authorship information and citations (scientific papers), people commenting or following the blog of another person, and so on. From that perspective, we would like to combine the two modes (i.e. textual content and social links) in order to leverage the performance of a general person retrieval system.

Depending on the application, the problem could take different forms, for instance: (1) document author prediction, (2) recipient recommendation and (3) alias detection. In the first case (i.e. author prediction), the query is a document with some social metadata but whose author is unknown. In the second case (i.e. recipient recommendation), the query still consists of a document whose authors (and possibly some recipients) are known and the goal is to extend the list of possible recipients. In the third case (i.e. alias detection), the query takes the form of a set of documents (with social metadata) attached to some person whose social role with respect to these documents could be author, recipient, follower, commenter, *etc*.

Traditional approaches tackle this problem by early or late fusion (see for instance (McDonald and Smeaton, 2005; Macskassy, 2007)). Typically, a person is represented by two sets of features: a word profile and a contact/link profile. These profiles could be distribution over words or over persons or any other vectorial representation, possibly using standard weighting schemes of information retrieval. Both profiles (textual and social) are typically obtained by aggregating the corresponding features of the docu-

Mantrach A. and Renders J..
PEOPLE RETRIEVAL LEVERAGING TEXTUAL AND SOCIAL DATA.
DOI: 10.5220/0003669203250333
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 325-333
ISBN: 978-989-8425-79-9

ments in which they play a role. If an early fusion approach is adopted, these two profiles are joined and relevance scores are then determined by computing some adequate similarity measure between the joined profile of a person and the query (that is supposed to contain both aspects as well). If a late fusion approach is used, similarity measures are computed for each mode separately and then merged by an aggregation operator such as a weighted average, a soft-min or a soft-max applied to normalized scores.

Recently, cross-modal similarity measures based on "trans-media relevance feedback" ((Clinchant et al., 2007; Mensink et al., 2010)) have been proposed for multimedia information retrieval, as an extension of the standard pseudo-relevance feedback mechanism in monomodal information retrieval. "Trans-media relevance feedback" could be considered as a two-step process, where the query is first enriched by parts of the documents of the target collection that are its nearest neighbors in one mode and where the second step determines the relevance scores as the similarity between the enriched query and the documents in the other mode. As far as we know, "trans-media relevance feedback" has never been applied in order to combine text with social metadata. So, inspired by these cross-modal similarity measures initially designed to combine image and text, we propose in this paper new ways of combining social and content aspects for retrieving people from a collection of documents with social metadata. To this aim, we first define a set of basic multimodal similarity measures between socially-labelled documents and queries, that could then be aggregated at the person level to provide a final relevance score for the general people retrieval problem. Then, we examine particular instances of this problem, namely author retrieval, recipient recommendation and alias detection, by describing in each case how the general method particularizes for the problem.

This paper is organized as follows. After this introduction, we formalize the main problem, as well as its particular instances, introducing the notations and the mono-modal similarity measures. Then, we explain a novel general approach taking into account mono-modal and cross-modal similarities and describe the particular algorithms for the different instances of the main problem (*i.e.* author prediction, recipient recommendation and alias detection). Later, we introduce the ENRON email collection and report the experimental results obtained on this data set for two tasks: author prediction and alias detection. Finally, after making some links with related works in a section dedicated to prior art, we conclude this paper by perspectives and future avenues to be explored.

2 PROBLEM STATEMENT

We consider the case of a collection of textual documents with social metadata, such that documents could be both indexed by words and by participants. Persons belong to the closed set $\mathcal{P}$, indexed by p which takes its value in $\{1,\ldots,P\}$, where P is the total number of participants involved in the documents. A participant could play different roles with respect to documents (author, recipient, commenter, follower, *etc.*). Roles belong to the closed set $\mathcal{R}$ and are indexed by a variable denoted as r (we suppose that there are R possible roles, so that r takes its value in $\{1,\ldots,R\}$). In this multi-view setting, a document could be represented by the traditional term-document matrix denoted by $\mathbf{X}$ (whose element $X(i,j)$ is the number of occurrences of word i in document j), but also by some matrices, denoted by $\mathbf{R}_r$, whose element $R_r(i,j)$ is 1 when the participant i plays role r in document j (and 0 otherwise). So, there are as many $\mathbf{R}_r$ matrices as there are possible roles. Sometimes, it could be convenient to not distinguish between any particular role in the social metadata; in this case, we represent the social information by a single $\mathbf{R}$ matrix whose element $R(i,j)$ is 1 when person i is a participant of document j.

Generally, the user submits a multi-faceted query and targets persons who are participant in the documents of the collection, possibly under some specific roles, called the target role(s). Formally, we will consider the general case where the query $\mathbf{Q}$ consists of a set of K sub-queries $\mathbf{q}_k$ where $k = 1,\ldots,K$. Each sub-query $\mathbf{q}_k$ is itself a triplet ($\mathbf{qx}_k$, $\mathbf{qr}^r_k$, r'_k) where $\mathbf{qx}_k$ is a vector whose element $qx_k(i)$ is the number of occurrences of term i in the k^{th} sub-query; $\mathbf{qr}^r_k$ is a vector whose element $qr^r_k(i)$ is 1 if person i plays a role r in the sub-query $\mathbf{q}_k$ (and 0 in the other cases); and r'_k is the target role of the sub-query $\mathbf{q}_k$. Note that we assumed that each sub-query gives key persons only for one role, but this can be easily extended so that sub-queries span multiple roles. So, the objective of the person retrieval problem is to rank each person of the socially-labelled collection with respect to their relevance to all multi-faceted sub-queries $\mathbf{q}_k$ ($k = 1,\ldots,K$), each of them associated with a target role r'_k.

Let us be more concrete by considering the task of author prediction of a new document (a scientific paper like this one). Persons participating in documents could have two roles: "is_author_of" or "is_cited_in". The query is, in this case, made of only one single sub-query (k=1) that has the following facets: the words of the new document ($\mathbf{qx}_k$) and the persons cited in this document ($\mathbf{qr}^r_k$, where r refers here only to the "is_cited_in" role); the target role r'_k is obviously

the "is_author_of" role. Moreover, let us consider now the task of recipient recommendation or, equivalently, the task of contextual citation proposal. The query is made of only one single sub-query that has at least the following facets: the words of the new document ($\mathbf{qx}_k$) and the author of the document ($\mathbf{qr}_k^r$, where r refers here only to the "is_author_of" role). If the new document already contains a partial list of citations (or recipients), the query has an extra facet which is the vector of cited persons. The target role r'_k is of course the "is_cited_in" role (or "is_interested_in").

The case of alias detection could also be formalized in this framework. Suppose that we have to decide whether a person p associated to a set of K documents in which she/he participated coincides with any other person in the socially-labelled collection. We could decide to simply index the base of documents by their standard textual content and by their participants (in other words, we merge the active and passive roles as a single, undirected role, called "participates_in"). The query now consists of the set of K documents where p was involved and each sub-query $\mathbf{q}_k$ ($k = 1, \ldots, K$) is a multi-faceted document with its textual content ($\mathbf{qx}_k$) and its participants ($\mathbf{qr}_k^r$, where r designates the "participates_in" role). The target roles r'_k are, in this case, identical to r, i.e. the "participates_in" role. Alternatively, we could make the distinction between the active and passive roles and tackle the alias prediction problem as the superposition of an author prediction task and a recipient prediction task. In this case, there will be subqueries containing key-persons playing an active role ("is_author_of" for instance) and subqueries with key-persons playing a passive role ("is_follower_of", "is_recipient_of", *etc.*); subqueries with active key-persons will be associated to a passive target role, while subqueries with passive key-persons will be associated to an active target role.

3 MULTI-STEP SIMILARITIES AND AGGREGATION

We are given a collection of socially-labelled documents represented by $\mathbf{X}$, $\mathbf{R}_r$, and a query $\mathbf{Q}$ represented by a set of triplets ($\mathbf{qx}_k$, $\mathbf{qr}_k^r$, r'_k) where the subscripts r and k designate role and sub-query indices respectively while r'_k is the target role of the k^{th} sub-query. We want to compute a relevance score $RSV(p, \mathbf{Q})$ that measures how well person p is relevant to query $\mathbf{Q}$. The method involves two phases. In the first phase, we compute the multi-modal similarities between the sub-queries and the documents of the socially-labelled collection. In the second phase, we aggregate these similarities in order to go from documents to persons (by a weighted average of the similarities with the documents for which the persons play the target role r'_k) and from sub-queries to query (simply by averaging over each sub-query).

For the first phase, we define two types of similarities. The "one-step" similarities are simply the standard mono-modal similarities. The "two-step" similarities implement the "trans-media" (or mono-media) pseudo-relevance feedback mechanisms that we mentioned before (see also (Clinchant et al., 2007; Mensink et al., 2010)). We start by defining the basic, mono-modal similarity measures commonly used to compute the similarity between a sub-query $\mathbf{qx}_k$ and a document d of the collection (or between two documents of the collection). The Language Modelling (LM) approach to Information Retrieval with Jelinek-Mercer smoothing ((Zhai and Lafferty, 2001)) served as basis in order to define these (asymmetric) similarity measures, and in particular the query log-likelihood criterion. When the textual content is considered, the similarity based on the query log-likelihood criterion could be computed as: $sim_T(\mathbf{q}_k, d) = \sum_w qx_k(w) \log\left(1 + \frac{\lambda_T}{1-\lambda_T} \frac{X(w,d)}{p(w).L(d)}\right)$, where w is an index over the words of the vocabulary, λ_T is the smoothing factor used in the Jelinek-Mercer LM smoothing strategy, $p(w)$ is the collection-based prior probability of word w (number of occurrences of w in the whole collection, divided by the total number of words in the collection) and $L(d)$ is the length (in words) of document d. Similarly, for a social role r, we could define the asymmetric similarity measure: $sim_r(\mathbf{q}_k, d) = \sum_p qr_k^r(p) \log\left(1 + \frac{\lambda_r}{1-\lambda_r} \frac{R_r(p,d)}{p_r(p).L_r(d)}\right)$, where p is an index over the participants, λ_r is the smoothing factor, $p_r(p)$ is the collection-based prior probability of person p in role r (number of times that p plays role r in the whole collection, divided by the total number of times that any person plays this role in the collection) and $L_r(d)$ is the number of persons playing role r in document d.

In this framework, we can define "two-step" similarity measures as well. For clarity of the presentation, we suppose that we have only one role r so that we have to combine only two views. But, of course, the same mechanisms could be used for dealing with all possible pairwise combinations of views. With two views, it is possible to define four different "two-step" similarity measures, namely $sim_{v_1,v_2}(\mathbf{q}_k, d) = \sum_{d' \in NN_{v_1}} sim_{v_1}(\mathbf{q}_k, d') sim_{v_2}(d', d)$ where v_1 and v_2 refer to one of the two modes (textual or social), NN_{v_1} designates the set of κ nearest neighbors of $\mathbf{q}_k$ using the mono-modal similarity measure in mode v_1 (typical values of κ are between 3 and 20 and could be different for each mode). Now, we can merge all these

similarity measures into one unique multi-modal similarity measure by a weighted average: $sim_G(\mathbf{q}_k, d) = \sum_v \alpha_v sim_v(\mathbf{q}_k, d) + \sum_{v_1,v_2} \alpha_{v_1,v_2} sim_{v_1,v_2}(\mathbf{q}_k, d)$ where v, v_1 and v_2 refer to the possible modes.

An obvious issue is how to choose these weights. We have adopted two extreme approaches. In the first one, we simply give an equal weight to each contribution, after studentization of the similarities (i.e. removing the mean of the scores over the documents and dividing the difference by the standard deviation). In the second one, we try to learn the optimal weights in order to maximize a utility function (typically the normalized discounted cumulated gain at rank 10, or NDCG@10) for a set of training queries with their corresponding relevance judgements. However, the learned α_i gave results very similar to the simple mean operator applied to the studentized scores. It seems that the optimum of this problem is quite "flat" and that it is not necessary to fine-tune the weights

Coming back to the second phase of the method, that is the aggregation phase, we now have to specify how we aggregate the subquery-document multimodal similarities into a final relevance score with respect to the query $\mathbf{Q}$ and the set of target roles r'_k. As there is no prior information giving more importance to a sub-query versus another, we can simply sum the contribution coming from each sub-query. Conversely, it could be interesting, when aggregating documents into person profiles that play the target role r'_k with respect to them, to weight the documents differently. Intuitively, when calculating a profile for a specific participant p, we might give less importance to documents in which more persons were involved with the role r'_k; using our previous notations, we could decide to weigh documents by the inverse of $L_{r'_k}(d)$. Note that, experimentally, using this weighting scheme always gave better performance than using the simple sum (or simple mean). So, finally, the aggregation equation of the second phase can be expressed as: $RSV(p, \mathbf{Q}) = \sum_k \sum_{d|p \in r'_k(d)} \frac{sim_G(\mathbf{q}_k, d)}{L_{r'_k}(d)}$, where $r'_k(d)$ denotes the set of persons playing the role r'_k in document d.

To be more concrete, we can synthesize the settings corresponding to particular instances of the general problem. For the author prediction task, the query consists of one document (K=1 :there is only one subquery) with textual content ($\mathbf{qx}$ vector) and social content ($\mathbf{qr}^r$ vector); the role r is "is_recipient_of" (directed). The target role r' is the "is_author_of" role. For the recipient recommendation task, the query consists of one document (K=1) with textual content ($\mathbf{qx}$ vector) and social content ($\mathbf{qr}^r$ vector); three choices are possible for the role r: either "is_author_of", and/or "is_recipient_of" (when we have already an incomplete list of recipients), or "participates_in" (undirected case). The target role r' is the "is_recipient_of" role. As far as the alias detection case is concerned, the query $\mathbf{Q}$ consists of K documents; each document, corresponding to each sub-query $\mathbf{q}_k$, has a textual content ($\mathbf{qx}_k$ vector) and a social content ($\mathbf{qr}^r_k$ vector); a simple choice for the role r (quite satisfying in practice) is the "participates_in" (undirected) role. The target roles r'_k are then also the "participates_in" roles. Alternatively, if we want to keep the active/passive distinction in the roles, we could distinguish sub-queries (documents) whose roles r are "is_recipient_of" and target roles are "is_author_of", from sub-queries whose roles r are "is_author_of" and target roles are "is_recipient_of".

4 EXPERIMENTS

4.1 The ENRON Data Set

This section describes the basic input dataset, that is common to all tasks. This dataset consists of a set of vectors and matrices that represent the whole ENRON corpus (Klimt and Yang, 2004), after linguistic preprocessing and metadata extraction. The linguistic preprocessing consists of removing some particular artefacts of the collection (for instance some recurrent footers, that have nothing to do with the original collection but indicate how the data were extracted), removing headers for emails (From/To/... fields), removing numerals and strings formed with non-alphanumeric characters, lowercasing all characters, removing stopwords as well as words occurring only once in the collection. There are two types of documents: documents are either (parent) emails or attachments (an attachment could be a spreadsheet, a power-point presentation, ...; the conversion to standard plain text is already given by the data provider). The ENRON collection contains 685,592 documents (455,449 are parent emails, 230,143 are attachments). We decided to process the attachments simply by merging their textual content with the content of the parent email, so that we have to deal only with parent emails. For parent emails, we have not only the content information, but also metadata. The metadata consist of: the custodian (i.e. the person who owns the mailbox from which this email is extracted), the date or timestamp, the Subject field (preprocessed in the same way as standard content text), the From field, the To field, the CC field. Note that the last three fields could be missing or empty.

4.2 Benchmark Protocol

Author Prediction. The goal here is to retrieve the email sender (i.e. the content of the "From" field) using all the available information (except, of course, the author himself). It is straightforward to note that the temporal aspect plays here a big role. Indeed, we suppose that it is easier to retrieve the authors of emails based on recent posts than on old ones. Hence, our training-test splits reflect this aspect. In this task we adopt the mailbox point of view : we consider only the emails coming from a specific user mailbox. It assumed that the emails of the collection are sorted by increasing order of their timestamp. We consider training sets made of 10% of the mailbox. After learning, the goal is to predict the author for the test set that corresponds to the next 10% emails in the temporal sequence. For example, when considering a training set made from the first 10% of the collection (i.e. in terms of timestamp), the test set consists then of the emails in the interval 10%-20%. By time-shifting the training and the test sets by 10%, we may consider training on the emails going from 10% to 20% and test on the next 10%, and so on. So that, finally, we define 9 possible training and test sets.

Alias Detection. For some specific participants, we want to find who are their aliases (if they are). To assess the performance in solving this task, we simulate the alias creation by splitting some participants into two identities: the original participant and a new (virtual) participant who has another person index. The goal is to be able to retrieve, for the original participants which have been split, the corresponding alias. For instance, for a specific participant, we keep its original identity in 20% of his emails (as sender and receiver) while switching his identity in the remaining 80% emails to the new identity (i.e. alias). This operation is done for 100 participants with different switching rates (20%, 40%, 60% and 80%). The 100 participants on which this operation is done are chosen at random between the people that are involved in at least 20 emails. This task makes more sense on a corporate point of view, or even on a wider set of emails where one want to detect any identity theft. Hence, this task has been assessed at the corporate level (i.e. on the whole `ENRON` data set).

Performance Measures. For both tasks, we measure the retrieval performance by the recall at rank 1 (R@1), at rank 3 (R@3), at rank 5 (R@5) and at rank 10 (R@10), knowing that, for each "query", there is only one single person who is really relevant. We measure also the Normalised Discounted Cumulative Gain, limited to rank 10 (NDCG@10) and on the whole set of persons (NDCG).

For the author prediction task, as the test set may contain emails whose author was not an author of the training set documents, we simply removed such emails when assessing the performance on the test sets.

4.3 Results and Discussions

Author Prediction. Two different mailboxes (`vince.j.kaminski@enron.com` and `tana.jones@enron.com`) are considered. These two mailboxes appear among the five mailboxes having the most emails in the data set. Performances are given for the algorithm synthesized in $RSV(p,\mathbf{Q})$, with different choices of the similarity function $sim_G(\mathbf{q}_k,d)$. Performances are averaged for the 9 training/test splits, and standard deviations are also given. The results are reported for the following similarity measures (see Table 1, Table 2) : (1) monomodal textual similarities ($sim_T(\mathbf{q}_k,d)$), (2) monomodal social similarities ($sim_r(\mathbf{q}_k,d)$, where r is the "is_recipient_of" role), (3) text-text (two-step) similarities ($sim_{T,T}(\mathbf{q}_k,d)$), (4) social-social (two-step) similarities ($sim_{r,r}(\mathbf{q}_k,d)$), (5) social-text (two-step) similarities ($sim_{r,T}(\mathbf{q}_k,d)$), (6) text-social (two-step) similarities ($sim_{T,r}(\mathbf{q}_k,d)$), (7) baseline which is the simple sum of (1) and (2) and, finally, (8) the weighted sum of the previous similarities (after score studentization for each type of similarity) using equal weights ($sim_G(\mathbf{q}_k,d)$). We observe that the combination of the different similarity measures allow to improve significantly all performances.

Depending on the mailbox, the dominant mode may change: sometimes, using purely textual similarities is better than using the social similarities; and sometimes, this is the opposite. It appears, logically, that the best enrichment strategy (first step of the two-step similarities) depends on the dominant mode. For instance, when considering `vince.j.kaminski@enron.com`'s mailbox (Table 1), the social similarity score ($sim_r(\mathbf{q}_k,d)$) is better and therefore enriching the content by the nearest neighbors in the social mode is preferable (see ($sim_{r,T}(\mathbf{q}_k,d) > sim_T(\mathbf{q}_k,d)$) on Table 1). Conversely, when the dominant mode is the textual one as for `tana.jones@enron.com`'s mailbox, then enriching the query trough that mode is a better strategy (Table 2).

Alias Detection. For alias detection, we have decided to consider only a single social role in the query and in the documents, namely the "participates_in" role. In other words, we assume that what matters in a document is the fact that this document links people,

Table 1: Averaged performance measures on the different time slots of size 10 % (i.e. training size of 10 %) on user vince.j.kaminski@enron.com's mailbox for the task of author prediction.

Similarity	R@1	R@3	R@5	R@10	NDCG@10	NDCG
$sim_T(\mathbf{q}_k,d)$	0.33 ± 0.07	0.39 ± 0.07	0.42 ± 0.07	0.46 ± 0.08	0.39 ± 0.07	0.45 ± 0.06
$sim_r(\mathbf{q}_k,d)$	0.42 ± 0.07	0.52 ± 0.08	0.57 ± 0.07	0.66 ± 0.07	0.53 ± 0.07	0.58 ± 0.05
$sim_{T,T}(\mathbf{q}_k,d)$	0.40 ± 0.09	0.48 ± 0.09	0.52 ± 0.10	0.58 ± 0.09	0.48 ± 0.09	0.54 ± 0.07
$sim_{r,r}(\mathbf{q}_k,d)$	0.29 ± 0.08	0.43 ± 0.07	0.47 ± 0.06	0.56 ± 0.06	0.42 ± 0.06	0.48 ± 0.05
$sim_{r,T}(\mathbf{q}_k,d)$	0.39 ± 0.09	0.47 ± 0.08	0.52 ± 0.10	0.57 ± 0.09	0.47 ± 0.09	0.53 ± 0.07
$sim_{T,r}(\mathbf{q}_k,d)$	0.27 ± 0.13	0.45 ± 0.10	0.51 ± 0.09	0.60 ± 0.07	0.43 ± 0.09	0.49 ± 0.08
baseline	0.47 ± 0.07	0.60 ± 0.08	0.65 ± 0.07	0.73 ± 0.06	0.59 ± 0.07	0.64 ± 0.05
$sim_G(\mathbf{q}_k,d)$	0.50 ± 0.09	0.64 ± 0.09	0.69 ± 0.08	0.77 ± 0.07	0.63 ± 0.08	0.67 ± 0.07

Table 2: Averaged performance measures on the different time slots of size 10 % (i.e. training size of 10 %) on user tana.jones@enron.com's mailbox for the task of author prediction.

Similarity	R@1	R@3	R@5	R@10	NDCG@10	NDCG
$sim_T(\mathbf{q}_k,d)$	0.42 ± 0.17	0.63 ± 0.09	0.65 ± 0.09	0.68 ± 0.09	0.57 ± 0.10	0.60 ± 0.09
$sim_r(\mathbf{q}_k,d)$	0.45 ± 0.14	0.51 ± 0.16	0.52 ± 0.17	0.56 ± 0.18	0.50 ± 0.16	0.55 ± 0.15
$sim_{T,T}(\mathbf{q}_k,d)$	0.43 ± 0.16	0.65 ± 0.04	0.68 ± 0.04	0.71 ± 0.04	0.59 ± 0.06	0.62 ± 0.07
$sim_{r,r}(\mathbf{q}_k,d)$	0.33 ± 0.21	0.47 ± 0.26	0.49 ± 0.27	0.52 ± 0.28	0.43 ± 0.25	0.48 ± 0.23
$sim_{r,T}(\mathbf{q}_k,d)$	0.30 ± 0.25	0.47 ± 0.25	0.50 ± 0.26	0.53 ± 0.27	0.42 ± 0.25	0.47 ± 0.23
$sim_{T,r}(\mathbf{q}_k,d)$	0.45 ± 0.13	0.57 ± 0.16	0.60 ± 0.16	0.62 ± 0.16	0.54 ± 0.15	0.59 ± 0.13
baseline	0.55 ± 0.09	0.74 ± 0.09	0.77 ± 0.04	0.80 ± 0.04	0.69 ± 0.04	0.72 ± 0.04
$sim_G(\mathbf{q}_k,d)$	0.58 ± 0.08	0.79 ± 0.05	0.81 ± 0.05	0.85 ± 0.05	0.73 ± 0.04	0.75 ± 0.04

independently of their active or passive roles. Results are given for a set of 100 "fake" persons that correspond to original persons who have been randomly split and switched to another identity with varying switch rates (from 20 % to 80 %). Performances are averaged over these 100 persons.

As in the author prediction task, the results are reported for the 6 types of similarity measures (textual, social, textual-textual, social-social, textual-social and social-textual) and for the combination obtained by simple average of the studentized scores.

For sake of completeness, we have also reported on the tables the performance of the alternative approach (denoted by $sim_{G_{\text{alternative}}}$), namely first aggregating the sub-queries (into a single global query that is the multi-faceted person profile of the candidate alias) and the documents (to build person profile for each person of the collection): the final relevance score is obtained by late fusion (with equal weights) of the similarities of the aggregated textual and social profiles.

Here, the social-based similarity measures seem to provide better results than the textual ones (see Table 3, 4, 5 and 6). As in the author prediction task, it appears that it is better to enrich the query (first step in the "two-step" similarities) by its nearest neighbors in the dominant mode. Indeed, we observe that query enrichment through social similarities is nearly always better than through textual similarities (i.e. performance with $sim_{r,T}(\mathbf{q}_k,d)$ is better than performance using $sim_{T,T}(\mathbf{q}_k,d)$; and $sim_{r,r}(\mathbf{q}_k,d)$ is better than $sim_{T,r}(\mathbf{q}_k,d)$). Again, the simple mean combination of studentized scores outperforms the results of any single similarity. There is also a significant gain in using the "compute document similarities and then aggregate" scheme over the "aggregate documents and then compute person similarities" scheme.

5 RELATED WORK

To solve the multi-view problem, a common approach is to work on a graph representation of the data. For instance, (Slattery and Mitchell, 2000) exploit hyperlinks between web pages in order to improve traditional classification tasks using only the content. (Joachims et al., 2001) studied the composition of kernels in order to improve the performance of a soft-margin support vector machine classifier. In the same spirit, (Cohn and Hofmann, 2000; Zhu et al., 2007) improved the classification performance by using a combination of link-based and content-based probabilistic models. (Fisher and Everson, 2003) showed that link information can be useful when the document collection has a sufficiently high link density and links are of sufficiently high quality. In the same context, (Chakrabarti et al., 1998; Oh et al., 2000) use both local text in a document as well as the distribution of the estimated classes of other documents in its neighborhood, to refine the class distribution of

Table 3: Performance measures for the alias detection task where for 100 randomly selected users, 80 percent of their original emails have been reattributed to a new participant.

Similarity	R@1	R@3	R@5	R@10	NDCG@10	NDCG
$sim_T(\mathbf{Q},d)$	0.330	0.490	0.530	0.600	0.463	0.530
$sim_r(\mathbf{Q},d)$	0.410	0.540	0.580	0.620	0.515	0.581
$sim_{T,T}(\mathbf{Q},d)$	0.300	0.410	0.440	0.510	0.401	0.479
$sim_{r,r}(\mathbf{Q},d)$	0.430	0.610	0.640	0.680	0.562	0.632
$sim_{r,T}(\mathbf{Q},d)$	0.410	0.570	0.600	0.660	0.533	0.600
$sim_{T,r}(\mathbf{Q},d)$	0.270	0.410	0.500	0.570	0.415	0.496
$baseline$	0.470	0.580	0.670	0.690	0.583	0.639
$sim_G(\mathbf{Q},d)$	0.450	0.620	0.680	0.730	0.594	0.643
$sim_{G_{\text{alternative}}}(\mathbf{Q},d)$	0.40	0.57	0.65	0.69	0.55	0.60

Table 4: Performance measures for the alias detection task where for 100 randomly selected users, 60 percent of their original emails have been reattributed to a new participant.

Similarity	R@1	R@3	R@5	R@10	NDCG@10	NDCG
$sim_T(\mathbf{Q},d)$	0.490	0.600	0.640	0.670	0.582	0.641
$sim_r(\mathbf{Q},d)$	0.510	0.570	0.590	0.670	0.578	0.637
$sim_{T,T}(\mathbf{Q},d)$	0.480	0.580	0.600	0.670	0.572	0.628
$sim_{r,r}(\mathbf{Q},d)$	0.580	0.690	0.730	0.790	0.682	0.726
$sim_{r,T}(\mathbf{Q},d)$	0.480	0.650	0.700	0.790	0.628	0.671
$sim_{T,r}(\mathbf{Q},d)$	0.440	0.580	0.620	0.700	0.563	0.620
$baseline$	0.550	0.640	0.710	0.740	0.642	0.692
$sim_G(\mathbf{Q},d)$	0.580	0.670	0.740	0.760	0.670	0.719
$sim_{G_{\text{alternative}}}(\mathbf{Q},d)$	0.54	0.64	0.68	0.73	0.63	0.68

Table 5: Performance measures for the alias detection task where, for 100 randomly selected users, 40 percent of their original emails have been reattributed to a new participant.

Similarity	R@1	R@3	R@5	R@10	NDCG@10	NDCG
$sim_T(\mathbf{Q},d)$	0.500	0.590	0.630	0.680	0.583	0.645
$sim_r(\mathbf{Q},d)$	0.510	0.610	0.670	0.710	0.607	0.658
$sim_{T,T}(\mathbf{Q},d)$	0.460	0.600	0.600	0.650	0.559	0.619
$sim_{r,r}(\mathbf{Q},d)$	0.560	0.680	0.750	0.800	0.675	0.718
$sim_{r,T}(\mathbf{Q},d)$	0.450	0.680	0.710	0.760	0.619	0.669
$sim_{T,r}(\mathbf{Q},d)$	0.460	0.590	0.670	0.700	0.580	0.637
$baseline$	0.550	0.670	0.690	0.740	0.646	0.695
$sim_G(\mathbf{Q},d)$	0.630	0.730	0.740	0.780	0.706	0.749
$sim_{G_{\text{alternative}}}(\mathbf{Q},d)$	0.58	0.67	0.70	0.75	0.66	0.70

Table 6: Performance measures for the alias detection task where for 100 randomly selected users, 20 percent of their original emails have been reattributed to a new participant.

Similarity	R@1	R@3	R@5	R@10	NDCG@10	NDCG
$sim_T(\mathbf{Q},d)$	0.450	0.570	0.620	0.690	0.564	0.623
$sim_r(\mathbf{Q},d)$	0.460	0.540	0.610	0.670	0.554	0.611
$sim_{T,T}(\mathbf{Q},d)$	0.460	0.540	0.550	0.600	0.527	0.599
$sim_{r,r}(\mathbf{Q},d)$	0.450	0.610	0.650	0.720	0.587	0.647
$sim_{r,T}(\mathbf{Q},d)$	0.430	0.610	0.650	0.700	0.567	0.628
$sim_{T,r}(\mathbf{Q},d)$	0.470	0.590	0.610	0.680	0.569	0.631
$baseline$	0.540	0.650	0.670	0.720	0.626	0.678
$sim_G(\mathbf{Q},d)$	0.580	0.680	0.690	0.750	0.662	0.709
$sim_{G_{\text{alternative}}}(\mathbf{Q},d)$	0.54	0.6	0.66	0.72	0.62	0.67

the document being classified. (Calado et al., 2003) analyzed several distinct linkage similarity measures and determined which ones provide the best results in predicting the category of a document. They also proposed a Bayesian network model that takes advantage of both the information provided by a content-based classifier and the information provided by the document link structure. (Zhou and Burges, 2007) extended their transductive learning framework by combining the laplacian defined on each view. Moreover, (Macskassy, 2007) proposes to merge an inferred network and the link network into one global network. Then, he applies to that network an iterative classification algorithm based on relation labeling described in (Macskassy and Provost, 2007), a baseline algorithm in semi-supervised classification. Another related algorithm, namely "stacked sequential learning" has been used in order to augment an arbitrary base learning so as to make it aware of the labels of connected examples. (Maes et al., 2009) extended this last algorithm in order to decrease an intrinsic bias due to the iterative classification process. For its part, (Tang et al., 2009) solves a multiple graph clustering problem where each graph is approximated by matrix factorization with a graph-specific factor and a factor common to all graphs. Finally, more recently, (Backstrom and Leskovec, 2011) proposes to learn the weights of a namely "supervised random walk" using both the information from the network structure and the attribute data. People retrieval, or expert finding, has also been intensively studied this last years. Recently, (McCallum et al., 2007) proposed to apply his successful Author-Recipient-Topic (ART) model to an expert retrieval task. Therefore, they extended the ART model to the Role-Author-Recipient-Topic model in order to represent explicitly people's roles. During the same period, (Mimno and McCallum, 2007) introduced yet another topic based model, namely, the Author-Persona-Topic model for the problem of matching papers with reviewers. This family of works try to find latent variables that explain topics and communities formation and, indirectly, use these latent variables to compute the similarities, what is completely different from our approach. More related to our work, (Balog et al., 2009) proposed to model the process of expert search by introducing a theoretical language modeling framework. More recently, (Smirnova and Balog, 2011) proposed to extend this model with a user-oriented aspect in order to balance the retained expert candidate with the time needed by the user to contact him. Actually, these frameworks are mono-modal (i.e. working only on document terms), they do not consider any social or link information. Moreover, there is no aspect of pseudo-relevance feedback in order to enrich the submitted query.

6 CONCLUSIONS

In this work we presented a global framework for people retrieval in a collection of socially-labelled documents, which extends the classical paradigm of document retrieval by focusing on people and social roles. This framework may be applied to a wide range of retrieval tasks involving multi-view aspects. Our approach consists of separating the problem into two phases : in the first one (at the document level), we define valuable similarity measures exploiting direct (i.e. one step) and indirect (i.e. two-step, as in traditional pseudo-relevance feedback) relations between the query and the targeted collection. By this way, we are also able to capture cross-modal similarities in order to improve the final ranking. It appears that combining these similarities by a simple mean after score studentization offers a performance level that more complex combination schemes (for instance, learning the combination weights by a logistic regression when we can formulate the task as a supervised prediction problem) are not able to beat.

REFERENCES

Backstrom, L. and Leskovec, J. (2011). Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China*, pages 635–644.

Balog, K., Azzopardi, L., and de Rijke, M. (2009). A language modeling framework for expert finding. *Inf. Process. Manage.*, 45(1):1–19.

Calado, P., Cristo, M., Moura, E., Ziviani, N., Ribeiro-Neto, B., and Gonçalves, M. (2003). Combining link-based and content-based methods for web document classification. In *Proceedings of the twelfth international Conference on Information and Knowledge Management (CIKM 2003*, pages 394–401. ACM.

Chakrabarti, S., Dom, B., and Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pages 307–318.

Clinchant, S., Renders, J.-M., and Csurka, G. (2007). Transmedia pseudo-relevance feedback methods in multimedia retrieval. In *CLEF*, pages 569–576.

Cohn, D. A. and Hofmann, T. (2000). The missing link - a probabilistic model of document content and hypertext connectivity. In *Neural Information Processing Systems conference (NIPS 2000)*, pages 430–436.

Fisher, M. and Everson, R. (2003). When are links useful? Experiments in text classification. In *Advances in information retrieval: proceedings of the 25th European*

Conference on Information Retrieval Research (ECIR 2003), pages 41–56. Springer Verlag.

Joachims, T., Cristianini, N., and Shawe-Taylor, J. (2001). Composite kernels for hypertext categorisation. In *Proceedings of the International Conference on Machine Learning (ICML 2001)*, pages 250–257.

Klimt, B. and Yang, Y. (2004). The enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning, Pisa, Italy, September 20-24*, pages 217–226.

Macskassy, S. A. (2007). Improving learning in networked data by combining explicit and mined links. In *Proceedings of the 22th conference on Artificial Intelligence (AAAI 2007)*, pages 590–595.

Macskassy, S. A. and Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983.

Maes, F., Peters, S., Denoyer, L., and Gallinari, P. (2009). Simulated iterative classification: a new learning procedure for graph labeling. In *Proceedings of the European Conference on Machine Learning (ECML 2009)*, pages 47–62.

McCallum, A., Wang, X., and Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Intell. Res. (JAIR)*, 30:249–272.

McDonald, K. and Smeaton, A. F. (2005). A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *Proceedings of 4th International Conference on Image and Video Retrieval, CIVR 2005, Singapore, July 20-22, 2005*, pages 61–70.

Mensink, T., Verbeek, J. J., and Csurka, G. (2010). Trans media relevance feedback for image autoannotation. In *Proceedings of British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010*, pages 1–12.

Mimno, D. M. and McCallum, A. (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 500–509.

Oh, H., Myaeng, S., and Lee, M. (2000). A practical hypertext catergorization method using links and incrementally available class information. In *Proceedings of the 23rd international ACM conference on Research and Development in Information Retrieval (SIGIR 2000)*, pages 264–271. ACM.

Slattery, S. and Mitchell, T. (2000). Discovering test set regularities in relational domains. In *Proceedings of the 7th international conference on Machine Learning (ICML 2000)*, pages 895–902.

Smirnova, E. and Balog, K. (2011). A user-oriented model for expert finding. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, pages 580–592.

Tang, W., Lu, Z., and Dhillon, I. S. (2009). Clustering with multiple graphs. In *Proceeding of The Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, 6-9 December 2009*, pages 1016–1021.

Zhai, C. and Lafferty, J. D. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 334–342.

Zhou, D. and Burges, C. J. C. (2007). Spectral clustering and transductive learning with multiple views. In *Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvalis, Oregon, USA, June 20-24, 2007*, pages 1159–1166.

Zhu, S., Yu, K., Chi, Y., and Gong, Y. (2007). Combining content and link for classification using matrix factorization. In *Proceedings of the 30th international ACM conference on Research and Development in Information Retrieval (SIGIR 2007)*, pages 487–494. ACM.