From Entities to Geometry: Towards exploiting Multiple Sources to Predict Relevance

Emanuele Di Buccio
Department of Information Engineering
University of Padua, Italy
dibuccio@dei.unipd.it

Mounia Lalmas
Department of Computing Science
University of Glasgow, UK
mounia@acm.org

Massimo Melucci
Department of Information Engineering
University of Padua, Italy
melo@dei.unipd.it

ABSTRACT

The goal of an Information Retrieval (IR) system is to predict which information objects can help users in satisfying their information needs, i.e. predict relevance. Different sources of evidence can be exploited for this purpose. These sources are the properties of the different entities involved when retrieving and accessing information, where examples of entities include the information objects, the task, the user, or the location. The main hypothesis of this paper is that, to exploit the variety of entities and sources, it is necessary to model the relationships existing between the entities and those existing between the properties of the entities. Such relationships are themselves possible sources that can be used to predict relevance. This paper proposes a methodology that supports the design of an IR system able to model in a uniform way the properties of the entities involved, the properties of their relationships and the relationships between the different properties. The methodology is structured in four steps, aiming, respectively, at supporting the selection of the sources, collecting the evidence, modeling the sources and their relationships, and using the latter two to predict relevance. Sources and relationships are modeled and then exploited through a previously proposed geometric framework, which provides a uniform and concrete representation in terms of vector subspaces.

1. INTRODUCTION

The goal of an IR system is to predict which information objects can help users in satisfying their information needs. For instance, if the information need is expressed by the user as a textual query, the IR system has to predict which documents are relevant to the formulated query. According to this interpretation, IR can be framed as a problem of evidence and prediction [1]. The prediction can be performed through the different sources of *evidence* involved in the retrieval process. Content, meta-data and annotations of the information objects are examples of such sources, and have been used by many retrieval systems.

These sources have been shown to be effective to predict relevance, but other sources exist. An example is the behavior of the user during the search process, for instance described in terms of interaction features – display time, click-through data, amount of scrolling, or other features e.g. [2]. These features have been adopted as sources of evidence to estimate relevance, e.g. display-time in [3], click-through data in [4], or a combination of several features in [5, 6]. Nowadays commercially available devices, e.g. mobile phones, are equipped with tools that can capture information about the user location and from the surrounding environment, besides having access to all the information provided by the web or the user personal data.

The various sources may not have the same impact in predicting relevance, and as such their relative contributions should be investigated. For instance ranking algorithms that are based on different object representations will usually return sets of relevant information objects with little overlap [8]. It is therefore important, as stated in [8], to "explicitly describe and combine multiple sources of evidence about relevance" when developing ranking algorithms. More precisely, it is important to explicitly consider the relationships existing between sources. However, the design and the implementation of distinct ranking algorithms, one for each type of sources, may not allow for considering relationships between sources. It is thus important to investigate approaches that combine evidences rather than approaches that combine ranking algorithms. This would allow for the relationships between sources to be explicitly integrated in the ranking algorithm.

This paper proposes a methodology that supports the design of an IR system able to model in a uniform way the properties of the entities involved, the properties of their relationships and the relationships between the different properties. The methodology is structured in four steps, aiming, respectively, at supporting the selection of the sources, collecting the evidence, modeling the sources and their relationships, and using the latter two to predict relevance. The last two steps are based on the geometric framework proposed in [9], which provides a uniform and concrete representation of the sources and their relationships in terms of vector subspaces.

The methodology aims at being general, in the sense that it is not related to a specific source or set of sources. However, for illustration purpose, two sources will be considered in this paper, namely, the content of the information objects to be ranked and the behavior of the users when accessing or retrieving information. The former has been selected because past research in IR provides a number of representations of the content that have been shown to lead to effective retrieval [8]. The latter has been extensively investigated in

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

Information Science (IS) and has in the last decade become a subject of investigation in IR. Indeed, experimental evaluation has shown how usage data stored in transaction logs [3, 4, 6, 10] or so-called interactive IR systems [11, 12] can effectively predict relevance. The use of the Entity-Relationship database model for describing IR objects was introduced in [13] for automatic hypertext construction purpose – this paper enlarges that view and connect the entities and relationship at the conceptual level to a mathematical model which provides a language at the logical level.

2. MOTIVATIONS AND METHODOLOGY RATIONALE

IR systems can exploit the evidence provided by different sources to improve retrieval effectiveness. In [8] the author considers several document representations and discusses approaches to combine the contribution provided by each representation. In [14] the inference network framework is adopted to combine link-based evidence with content-based evidence for web retrieval. Evidence on the structure of the documents can be incorporated, for instance, using the Dempster-Shafer theory of evidence [15]. However, the different document representations are only a subset of the available sources.

Let us consider, for instance, the scenario where a user is looking for information about restaurants in London. If Venice is the location where the search is performed, this probably suggests that the user is planning a trip in London, and restaurants in an arbitrary London area may be of interest. If the search is performed on a mobile phone and the GPS position indicates that the user is in London, probably the user is more interested in restaurants near his current position. We can see that in this scenario, other units besides the information objects are involved. In this paper, we refer to units as *entities*. For instance, in our scenario, the entities involved are the user, the location, the task the user is performing when looking for information – i.e. "travel in London" – and the specific topic within the task[1] – i.e. "finding restaurants in London".

Each entity is characterized by a number of properties. When the entity is an "information object", examples of properties include content, meta-data and annotation. For the entity "location", instances of properties are the GPS position or the IP address.

Each entity exists independently of the properties we can observe about it, but the observed properties are the evidence that can be used to build a model of the entity, that is to obtain a description of the entity – in this work a mathematical description – that can be used to predict relevance. In other words, *the properties of the entities are the sources of evidence that can be exploited to help predicting the relevance of information objects.*

Not only the properties of the entities are sources of evidence, but also the relationships between entities (if any) can provide additional evidence to predict relevance. Let us consider a list of results returned by an IR system in response to a query and the user who formulated the query. The behavior of the user when examining a result is one of the properties to describe the relationship between the entity user and the entity result; such property constitutes a source that can be exploited to predict relevance. Indeed, research in Interactive IR has shown that a retrieval system can benefit from evidence gathered from the information seeking activities of a user. For example, Implicit Relevance Feedback (IRF) algorithms [10] exploit the information gathered from the interactions between the user and the documents to recommend query expansion terms or to re-rank documents. Even the concept of *relevance* can be defined as "a relation between a document and a person, relative to a given information need" [1], the document and the person being two entities.

The set of entities and relationships, and their properties, are neither fixed nor unique, as they depend on the specific retrieval application – e.g. the entity location is crucial for search carried out on a mobile phone or to customize search results according to the country where the search originates. Therefore, the *selection of the sources* is an important issue that needs to be addressed.

Once the appropriate sources have been identified, each of them has to be modeled, so that to be exploited for retrieval. In this work, we refer to the model of a source as a *dimension*. A first step to obtain a dimension is to identify a set of features that describe it. *Feature* here refers to the information obtained by the observation of a property of an entity or a relationship. For an entity "location" described by the dimension "GPS position", the features are the GPS position components. For a "web result" entity, the keywords in the title, the snippet or the URL of the result are example of features. Since the features constitute the evidence that model a source, a procedure to *select and collect features* has to be designed and implemented.

The description (model) of the sources is what get used to predict relevance. In this work the framework adopted to build the description is the vector subspace formalism proposed in [9]. The basic rationale for this is that we want to map the collected data, prepared in a matrix, in a new vector space basis – the vector subspace spanned by the basis is the model of the source.

Once a representation in terms of subspaces has been built both for the sources and the information objects, a trace-based function, the one exploited in [9], can be adopted to rank information objects by exploiting the information about the different sources of evidence that have been modeled. In other words the trace-based function, which we briefly describe in Section 4, is a tool to handle the *prediction* problem.

In summary, four steps have been identified, and each of them needs to be addressed to be able to predict relevance using multiple sources of evidence, namely, sources selection, features collection, source modeling and relevance prediction. Figure 1 illustrates these four steps for the relationship between the entities "user" and "information objects"; here, the relationship is characterized by the source "user behavior" described in terms of "interaction features".

In this paper we will focus on two of the above steps, specifically evidence collection and source modelling, which will be discussed respectively in Section 3 and Section 4; some remarks on the implementation of these methodology steps and their evaluation are reported in Section 5.

[1]We take the definition of *task* and *topic* from [2]: "Task was defined for this study as the goal of information-seeking behavior, and topic was defined as the specific subject within a task."

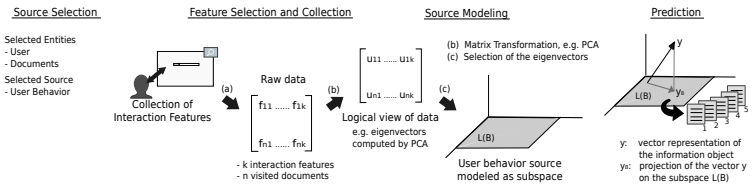

Figure 1: Methodology steps and specific application to the user interaction behavior.

3. EVIDENCE COLLECTION

Let us return to the scenario of a user looking for information about restaurants in London. Let us suppose the user, to satisfy his information need, interacts with a search engine and submits the query "restaurants in London". The search engine returns a ranked list of results. For simplicity, we focus on two entities only, namely, the user and the result. When examining the returned results, the user interacts with them and with the information objects the results refer to. In this scenario the behavior of the user when examining and (eventually) accessing the results can be considered as a property to describe the involved entities and, particularly, as a source to assist relevance prediction. In the above scenario another source available is the content of the abstracts (title, snippet and URL) of the results and the content of the corresponding information objects.

Once the sources have been selected, the next step is to *collect the evidence* to build the model of these sources. This step consists of selecting the features to be gathered to build a model of these sources, and then the actual collection of the selected features.

In the event of the source "user behavior" a possible choice, as depicted in step two of Figure 1, is the adoption of so-called interaction features. This is for instance the approach adopted in [5, 6] where several interaction features are exploited simultaneously. In particular, in [6] a subset of the features gathered in the user study described in [2] was exploited to obtain a vector subspace representation of the user behavior. When using a representation personalized for each user and tailored on the specific search task to re-rank the documents, the keywords extracted from the top re-ranked documents were shown to be effective as source for query expansion. The methodology proposed in that work assumed that the interaction features were available for all the documents to be re-ranked. But this assumption does not hold in our considered scenario, unless the documents have been already visited with regard to past queries when performing the same task. Therefore, the *availability* of the interaction features is an issue to address. A possible solution is not to consider the features with regard to a single user, but with regard to a group of users, e.g. performing the same task.

Another reason to exploit group interaction data is the *reliability* of the interaction features. The features need to be reliable indicators of the user needs, interests or intents. To clarify what we mean by "reliable feature", let us consider the display-time: this feature, when considered in isolation and referring to a single user, is subject to variations. Exploiting this feature when predicting relevance may be difficult [3], thus making it not reliable. But in [3] the authors found that display-time, when used as implicit measure, is more consistent when referring to multiple subjects performing the same task, than when personalized to each user.

Individual users and user groups, does not necessarily need to be considered as mutually exclusive sources for interaction features. For instance, in [5] user behavior models to predict user preferences for web ranking are learned by exploiting simultaneously feature values derived from the individual's behavior and those aggregated across all the users and search session for each query-URL pair.

The selection of the features of a source to then be gathered affects the modeling step, since they constitute the evidence used to build a model of the source. However, the procedure to collect features is part of the design of the IR system, in particular, the components aimed at gathering the selected features and managing them. For instance, when interaction features have been selected as implicit indicators, a browser extension can be used to monitor the gathering of such features. This is the approach adopted in the Lemur Query Log Project[2], a study to gather the query logs from users of the Lemur Query Log Toolbar[3][4]. It should be noted that the development of an extension that stores the usage data on the client side may encourage the user to adopt this monitoring tool since no personal data need to be provided to the server.

4. SOURCE MODELING AND PREDICTION

Once the evidence has been gathered, the next step consists of modeling the evidence so that it can be used to predict relevance. In this work the mathematical construct of the vector subspace is used for this purpose.

In this paper, the evidence gathered by the different sources is exploited to rank information objects with respect to a given query. This is done by using the different representations of the objects generated from the sources. For instance, if the user "interaction behavior" is a considered source, an information object can be described in terms of the interaction features monitored when a user is visiting the object — e.g. an object being displayed for 30 seconds, clicked 3 times and on which 5 scrolling actions have been performed, can be represented as the vector $y = (30, 3, 5)$. The same object, if the source "content" is considered, can be described as the vector of the TF·IDF weights of the terms appearing in it. The construct of the vector space basis is particularly suitable to model these multiple representations. Indeed, intuitively, the same information object can be represented with regard to different sources in the same way the same vector can be generated by different vector space basis.

A second reason to adopt the construct of the vector space basis is that some of the vector subspace representations

[2] `http://lemurstudy.cs.umass.edu/`
[3] `http://www.lemurproject.org/querylogtoolbar/`
[4] The goal of the study is to create a database of web search activities that will be provided to the information retrieval research community.

may reveal the logical structure underlying the collected evidence. The collected data, prepared in a matrix, is a vector representation of the source. This data often may be noisy. A matrix transformation, namely a change of basis, can be applied to map the original view of the data to one that is less noisy. Let us consider the re-evaluation of the Vector Space Model (VSM) proposed in [16]. The authors point out how some assumptions underlying the traditional VSM [17] – e.g. that the terms are orthogonal – may suggest that the vector was interpreted as a data structure and not as a logical construct. Subsequent developments show how the vector can be used as a logical construct able to capture dependencies between terms and between documents [16, 18]. The "latent semantics" [18] of the terms in the documents, that is the dependencies between terms, was used as a source for implementing a Pseudo Relevance Feedback algorithm [9] and an Explicit Relevance Feedback algorithm [19] based on the geometric framework adopted in this work.

To explain the role of the matrix transformation techniques in the modeling step, we use the example of information behavior as a source, where the latter is described in terms of interaction features. A matrix A can be prepared where the element (i, j) is feature j observed during the visit of object i, e.g. a display-time of 30 seconds. The matrix A, as mentioned above, can be a noisy vector-based representation of the observed data. A matrix transformation technique such as Principal Component Analysis (PCA) of $A^T A$ can be used to compute a new vector space basis – this is actually the approach proposed in [6]. PCA provides a set of eigenvectors and a subset of them can be used to obtain the user interaction behavior dimension – the model of the source is the subspace spanned by the eigenvectors. As suggested by this example, this geometric framework allows us to achieve one of our goals, which is to generate a representation of the properties of the relationships between entities – in the example mentioned above the user behavior was the property to be modeled.

The two mentioned approaches, that is the one adopted in [9, 19] and that adopted in [6], provide a solution for two distinct sources. In the former case the modeled source is a property of an entity, namely the latent semantics of the terms in the documents. In the latter case, the modeled source is a property of a relationships between entities, namely the user interaction behavior. However, we are also interested in modeling relationships (if any) existing between the properties of the entities, namely between sources, e.g. between the latent semantics of the terms and the user interaction behavior – this is different from modeling properties of relationships, e.g. the user interaction behavior.

Let us return to the scenario of a user looking for information about restaurants in London and suppose the term "jazz" appears in the abstract of one of the displayed results. The user when examining the result may realize that he is more interested in jazz restaurants than in general ones. This example also emphasizes how different sources are not necessarily independent from each other. Indeed, the features observed for a source (e.g. the user behavior) can be "entangled" with the features observed for another source (e.g. the particular meaning of a query feature in the selected results).

The design of one approach per source may not be able to model relationships that may occur between sources and consequently to exploit them, as reported in [20]. In this work, we consider that the relationships are themselves sources. Therefore, it is better to not consider distinct mappings, one for each source, but to compute a single vector space basis to represent the relationships between sources.

The model of the sources can be used in the retrieval process once the information objects have been represented by the features selected to describe the sources. Indeed, the measure of the degree to which the modeled source occurs in an information object can be computed as the distance between the vector representation of the information object, which corresponds to a one-dimensional subspace, and the subspace modeling the source(s) spanned by the vector space basis computed in the source modeling step. This motivates the function proposed in [9], where the author showed how such function can be interpreted as a trace-based function and that the measure is a probability measure. The idea of using trace in IR, and in particular the density operators, was originally introduced in [21], and one of its important consequence – subsequently exploited in [9] – was to "establish a link between geometry and probability in vector spaces" [21].

5. IMPLEMENTATION AND EVALUATION

The specific implementation we are investigating concern the two mentioned sources, that is, the user behavior and the latent semantic of the terms in the information objects.

With respect to user behavior, we are focusing on two issues. The first is the selection of the source for interaction features since, as discussed in Section 3, both individual and user groups interaction data can be exploited to prepare the matrix A and to build the source model. In particular, we are investigating the difference between the two contributions in terms of retrieval effectiveness when PCA is adopted as the matrix transformation technique. PCA allows handing dimensionality reduction and capturing the relationships among the features in an unsupervised manner. However, as stated in [6], the problem is that the eigenvector whose components best combine the interaction features, is not necessarily the first principal eigenvector, and the best performance are achieved when the eigenvector is manually selected. For this reason we are investigating other unsupervised methods to obtain a vector subspace representation of the interaction data.

With respect to the latent semantics of terms, one issue under investigation is the selection of the terms in the feedback documents. Indeed, if the terms appearing in these documents are adopted as evidence to build a source model, one issue, particularly when real-time feedback is required, is to handle matrices whose dimensions are the number of distinct terms in the feedback documents. In this case a possible solution is the selection of a subset of the terms, e.g. the top weighted ones. However, this strategy has been shown to not be effective [19]; therefore, we are investigating selection criteria for "good terms".

Since the main objective of the methodology is to model relationships, we will look into the *relationships* between sources, and investigate their implementation using the proposed geometric framework, and their impact on retrieval effectiveness. Two approaches are possible. The first approach is to rank information objects separately according to different dimensions and then combine the rankings into one. The second approach is to model all the sources as a unique vector subspace and then rank the information ob-

jects against such subspace. The latter approach has the advantage of exploiting all the dimensions simultaneously, thus avoiding any loss of information that may arise from not considering relationships between sources (which is the case with the first approach). In particular, as for the user behavior source, we are investigating unsupervised approaches to model relationships among sources.

Evaluation is crucial to validate the implementation of the methodology. The main problem is the availability of datasets where information about user interaction behavior, the content of results and information objects are available. Transaction logs [7] can provide this data, but no explicit relevance judgments are available to validate the effectiveness of the approaches under investigation; existing datasets with this information are not publicly available.

6. CONCLUDING REMARKS

The purpose of this work was the introduction of a methodology that aims at exploiting evidence coming from multiple sources to predict the relevance of information objects for given queries. Four methodological steps are required to achieve this goal, namely, sources selection, features collection, dimension modeling and relevance prediction. The geometric framework proposed in [9] was chosen to implement the last two steps because it provides a uniform model for the sources, which can be used by to rank objects according to their estimated relevance.

Moreover, we discussed some issues to be addressed when implementing the methodology for two specific sources, that is the user interaction behavior and the latent semantic of the terms in the information objects. The issues specifically concern the evidence collection and source modeling steps.

In future work we want to further investigate the concepts adopted in this paper, namely, *entity*, *relationship*, *dimension* and *feature*. We chose these concepts as they relate to the view of the world to be modeled – in our case in order to predict relevance – which consists of entities and relationships, where the entities exists independently of their properties. The properties, namely the sources, are the information that can be obtained by the observation of entities and relationships between them. This is the same view of the world adopted in the Entity-Relationship (ER) model [22], the most widely used data model for the conceptual design of databases. In the ER model the result of the observation is a *value* and the mapping from the entities set (or the relationship set) to the value set is named *attribute*. The notion of feature adopted in this work can be compared to the ER notion of value set. Moreover the notion of dimension can be compared to the notion of attribute, since both refers to properties of entities and relationships.

The above discussion suggests investigate the relationships among the ER model, the geometric framework proposed in [9] and the methodology proposed in this paper.

Acknowledgements This research is partly funded by a Royal Society International Joint Project (2008/R4). Mounia Lalmas is currently funded by Microsoft Research/Royal Academy of Engineering.

7. REFERENCES

[1] S. E. Robertson, M. E. Maron, and W. S. Cooper. Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Development*, 1(1):1–21, 1982.

[2] D. Kelly. *Understanding implicit feedback and document preference: a naturalistic user study*. PhD thesis, New Brunswick, NJ, USA, 2004.

[3] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of CIKM'06*, pages 297–306, New York, NY, USA, 2006. ACM.

[4] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of KDD '02*, pages 133–142, New York, NY, USA, 2002. ACM.

[5] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of SIGIR '06*, pages 3–10, New York, NY, USA, 2006. ACM.

[6] M. Melucci and R.W. White. Utilizing a geometry of context for enhanced implicit feedback. In *Proceedings of CIKM'07*, pages 273–282, Lisbon, Portugal, 2007.

[7] B. Jansen. Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, 28(3):407–432, 2006.

[8] W.B. Croft. Combining approaches to information retrieval. *Advances in information retrieval*, 7:1–36, 2000.

[9] M. Melucci. A basis for information retrieval in context. *ACM TOIS*, 26(3):1–41, 2008.

[10] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[11] R.W. White, J.M. Jose, and I. Ruthven. An implicit feedback approach for interactive information retrieval. *IP&M*, 42(1):166–190, 2006.

[12] N. Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, 2008.

[13] M. Agosti, and F. Crestani A methodology for the automatic construction of a hypertext for information retrieval *Proc. of ACM SAC*, 745–753, Indianapolis, Indiana, United States, 1993.

[14] T. Tsikrika and M. Lalmas. Combining evidence for web retrieval using the inference network model: an experimental study. *IP&M*, 40(5):751–772, 2004.

[15] M. Lalmas and I. Ruthven. Representing and retrieving structured documents using the Dempster-Shafer theory of evidence: modelling and evaluation. *Journal of Documentation*, 54:529–565, 1998.

[16] S. K. M. Wong and V. V. Raghavan. Vector space model of information retrieval: a reevaluation. In *Proc. of SIGIR '84*, pages 167–185, Swinton, UK, 1984. British Computer Society.

[17] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[18] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JASIS*, 41:391–407, 1990.

[19] E. Di Buccio and M. Melucci. University of Padua at TREC 2009: Relevance Feedback Track. In *Proc. of TREC 2009*, Washington, DC, USA, 2009. To Appear.

[20] E. Di Buccio and M. Melucci. Towards a Methodology for Contextual Information Retrieval. In *Proc. of CIRSE 2009*, Tolouse, France, 2009.

[21] C.J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, New York, NY, USA, 2004.

[22] P.P. Chen. The entity-relationship model—toward a unified view of data. *ACM TODS*, 1(1):9–36, 1976.