Utilizing a Geometry of Context for Enhanced Implicit Feedback

Massimo Melucci
University of Padua
Department of Information Engineering
Via Gradenigo 6 — 35131 – Padova, Italy
melo@dei.unipd.it

Ryen W. White
Microsoft Research
One Microsoft Way
Redmond, WA 90852 USA
ryenw@microsoft.com

ABSTRACT

Implicit feedback algorithms utilize interaction between searchers and search systems to learn more about users' needs and interests than expressed in query statements alone. This additional information can be used to formulate improved queries or directly improve retrieval performance. In this paper we present a geometric framework that utilizes multiple sources of evidence present in this interaction context (e.g., display time, document retention) to develop enhanced implicit feedback models personalized for each user and tailored for each search task. We use rich interaction logs (and associated metadata such as relevance judgments), gathered during a longitudinal user study, as relevance stimuli to compare an implicit feedback algorithm developed using the framework with alternative algorithms. Our findings demonstrate both the effectiveness of our proposed algorithm and the potential value of incorporating multiple sources of interaction evidence when developing implicit feedback algorithms.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: relevance feedback, search process, retrieval model.

General Terms: Theory, Experimentation, Human Factors.

Keywords: Implicit relevance feedback, geometry of information retrieval.

1. INTRODUCTION

The effective use of most Information Retrieval (IR) systems requires people to be able to express their information needs as concise textual query statements. However, searchers can struggle to select the terms that will lead to the most effective retrieval, in particular if their information needs are vague [1] or their knowledge of system vocabulary or indexing is poor [2]. To address these shortcomings systems that use Relevance Feedback (RF) [3] can leverage explicit indications of user interest to circumvent problems in the direct specification of needs as queries, and drive an iterative search process that can support more effective retrieval [4]. Despite the promise of RF, users are reluctant to provide explicit feedback, generally because they do not understand its benefits or do not perceive it as being relevant to the attainment of their information goals [5]. Given the potential usefulness of RF it is prudent to research other ways in which relevance information can be gathered from users at minimal cost to them in terms of time or cognitive resources. One way to do this is to use the contextual information generated during the interaction between the user and information as implicit RF (IRF) [6], where visited documents to which certain relevance criteria apply are assumed to be relevant. Contextual features such as document display time (i.e., the amount of time a document is in focus in the Web browser or on the desktop), document retention (e.g., saving, printing), and document interaction (e.g., scrolling, click-through) can be mined and used as the basis for relevance criteria in IRF algorithms. These algorithms can suggest query expansion terms, retrieve new search results, or dynamically reorder existing results.

For simplicity, IRF algorithms traditionally use just one implicit feature as relevance criteria. The two most common features used are document display time or document visitation, both available for every document the user examines.[1] However, the use of these features as a relevance indicator is potentially problematic since there is mixed opinion about whether display time is an accurate predictor of relevance [7, 8], and whether visiting a document implies relevance [9, 10]. In addition, it has been shown through user experimentation that a single feature can vary greatly between users and search tasks [11]. Although this makes a strong case for personalization [12] it also means that implicit evidence can be unreliable as there are usually only a small number of relevant documents available for each user, each task, and each user/task pair. If we could capitalize on multiple aspects of user interaction substantially more evidence about preferences becomes obtainable, and more robust IRF algorithms could potentially be created [13].

In this paper we present a formal framework based on vector spaces that captures multiple aspects of user interaction and allows a new mathematical model of IRF to be developed. It uses display time, document retention, and interaction events to build a multi-faceted user interest profile. Since it uses more than one feature, IRF algorithms developed using the framework are less susceptible to feature bias

[1]This is not the case for other features such as document retention.

CIKM'07, November 6–8, 2007, Lisboa, Portugal.

than those using a single feature. We compare the retrieval performance of an IRF algorithm generated using the framework against RF baselines that use pseudo-relevance feedback (i.e., assume top search results are relevant), and the centroid of visited documents' interaction features to create a single interaction feature. As we will describe, combining multiple features results in an algorithm that leads to higher retrieval effectiveness.

The remainder of this paper is structured as follows. In Section 2 we describe related work on implicit feedback and vector spaces in IR. In Sections 3 and 4 we describe the formal framework used to model interaction context and its direct application in the work reported here. In Section 5 we describe a comparative experiment to test the effectiveness of our algorithm against alternatives. We present our findings in Section 6, discuss them in Section 7, and conclude in Section 8.

2. RELATED WORK

Interest in IRF to support information-seeking has grown in recent years given the attractiveness of being able to reduce the user burden of explicitly providing feedback whilst still being able to proactively support their search activities. Much of the research in this area has focused on the substitutability of IRF for explicit RF [13, 14], or the impact of task and user information on the reliability of interaction features [15, 16]. Fox et al. [13] showed that implicit ratings such as session duration and number of result sets returned can be indicative of user satisfaction. They also showed that the combination of several implicit features, including reading time and the way the user exited from the result page, can predict search result relevance. Joachims et al. [14] analyzed users' decision processes during search-result click-through and compared implicit feedback against manual relevance judgments. They concluded that although the interpretation of clicks as absolute relevance judgments is difficult, relative preferences derived from clicks are reasonably accurate. Kelly and Belkin [15] reported that display time was not indicative of document relevance, and that display times differ significantly according to specific task, and according to a specific user. White et al. [16] showed that factors such as task, user experience, and stage in the search can affect the potential usefulness of IRF. These studies are important in informing our understanding about IRF's potential, but they do not put these findings into practice in realistic settings.

Some attempts have been made to use IRF in search applications. Morita and Shinoda [7] explored how behaviors exhibited by users while reading articles from newsgroups could be used as IRF for profile acquisition and filtering. Budzik and Hammond [17] developed a system capable of automatically retrieving documents and recommending URLs to the user based on what the user was typing in a non-search application. White et al. [18] used reading time as a technique for automatically re-ranking sentence-based summaries for retrieved documents. Additional research in areas such as user modeling [19] and attentive systems [20, 21] has mainly focused on the use of IRF to infer the interests of an individual user based on that user's interactions, and typically restricts the source of IRF to a single behavior such as document display time, editing, or visitation. Related work by Horvitz et al. [22] employed multiple aspects of user interaction behavior but did not do so for IRF and not in the search domain. The IRF algorithm we propose in this paper supports the individual user in a similar way to many of these applications but uses multiple aspects of user interaction behavior.

The use of aggregate click-through statistics has recently emerged as an alternative to personalization as an application of IRF. Joachims [23] and Agichtein et al. [10] improved search engine ranking using the collective result click-through behavior of many Web searchers. Radlinksi and Joachims [24] augmented click-through data with additional evidence of query reformulation behavior. White et al. [25] used the browsing behavior of many users to direct individuals to authoritative resources for the query topic. These approaches may help all users in some way but do not directly help the individual user specify their needs more effectively; this is the aim of the framework we describe.

As can be seen above, most of the research on IRF has focused on empirical user studies where the behavior of the user is observed when interacting with the system. In contrast, this paper presents a geometric framework based on vector spaces that utilizes multiple sources of evidence present in this interaction context (e.g., display time, document retention) to develop enhanced implicit feedback models personalized for each user and tailored to each search task.

The use of vector spaces for modeling dates back to the early days of IR. The Vector-Space Model (VSM) gives an intuitive yet formal view of indexing and retrieval. The VSM has attracted many researchers and newcomers since when it was introduced in [26] and [27]. It has proved a very effective, sound framework in retrieving documents in different languages, on different subjects, of different sizes, and of different media, thanks to a number of proposed and tested weighting schemes and applications.

However, the VSM has often been seen as a "spreadsheet"-based way for designing systems. The use of matrices is due to the need to describe documents as tuples of word frequencies; the potential of vectors was limited by this view of the model. As a consequence, the potential of vector spaces for retrieval has not fully been exploited in practice, even though some attempts have been made in the past with some success. For example, an early effort to reevaluate the VSM is reported in [28]. The discussion on the assumptions and the potential is the subject of [29].

A recent reconsideration of the Geometry of IR was presented in [30]. In that book Hilbert's vector spaces are used to see documents as vectors, relevance as a linear transformation, relevance statuses as the eigenvalues of the linear transformation, and the computation of the probability of relevance of a document as the projection of the document vector onto an eigenvector of the linear operator. In other words, the size of the projection of the document vector onto an eigenvector of the operator is the probability that the document is about the relevance state represented by the corresponding eigenvalue.

Despite its apparent simplicity, the mathematical properties of vector spaces can be used to achieve good retrieval performance. In particular, the idea of using a *basis* of a vector space to represent context was proposed in [31, 32] and further investigated in [33] with regard to the extension towards Quantum Mechanics. In this paper these constructs are leveraged for addressing the issue of IRF and taking a step toward context-aware IR modeling.

The approach we adopt uses feature correlation as basic information for designing an IRF algorithm. The hypothesis that the feature-document frequency matrix contains information about the correlation among features and among documents was cited in [34], stated in [28] and was further exploited in [35] in defining Latent Semantic Indexing (LSI). The latter is a technique based on Singular Value Decomposition (SVD) which aims to decompose the correlation matrix and disclose the principal components used to represent fewer independent concepts than many inter-dependent variables. In this paper, the features are observed from user interaction and the contextual factors extracted through decomposition describe the factors which can be exploited for the IRF algorithm.

The next section describes how context is represented in our proposal.

3. THE GEOMETRY OF CONTEXT

Information-seeking activities do not occur in isolation from surrounding environmental and situational factors. On the contrary, context affects information seeking and retrieval. Therefore, it is little surprising that the use of contextual information has emerged as an area of great interest in IR and information-seeking research [36]. In this section we present a mathematical framework that can be used to represent this context in a way that can be leveraged by IR systems. In general, a mathematical framework plays the role of a formal description of an IR system that is necessary for arriving at algorithms and data structures that can effectively be implemented. In particular, vector spaces are used for defining the mathematical framework proposed in this paper because they have the properties for uniformly describing different IR frameworks [30] which may make it appropriate for modelling context over other mathematical models.

The features characterizing users, time, places, or anything emerging from user-system interaction form a notion which can be referred to as context. Since context is an ambiguous concept, the following definitions used in this paper are provided:

- Variable: refers to either an entity of the context, for example, user, task, topic, or document, or a relationship between entities, for example, relevance or aboutness.

- Dimension: refers to a property of an entity, for example, user behavior, task difficulty, topic clarity, document genre, or relevance.

- Factor: refers to a value of a property, for example, browsing, complex search task, difficult topic, relevant, non-relevant, or mathematical document.

These definitions are sufficiently precise for describing in this section a framework that can be used in algorithms automated by IR systems to support searching in context. Through their interaction behavior users can define their own interaction context that may be representative of contextual influence from the task or other constraints on them. For example, many short document display times may be indicative of a user attempting a time-constrained task. When some evidence is gathered from context, IRF can be performed for expanding queries, reordering retrieval results,

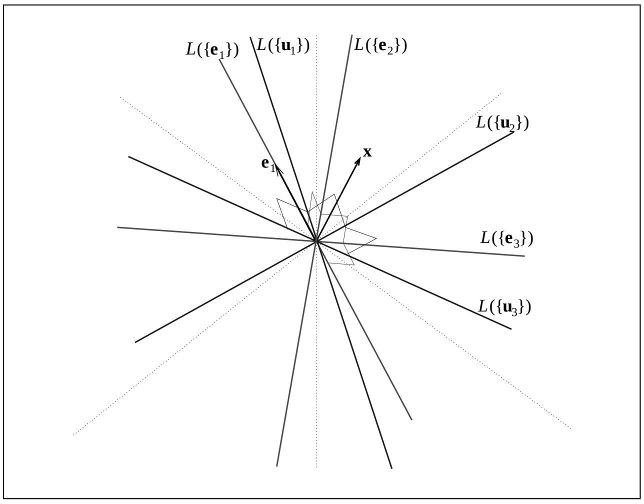

Figure 1: A vector is generated by infinite dimensions.

or re-searching. The framework described is based on Linear Algebra concepts which gives an algebraic representation of vectors, operators, and subspaces. As distances between vectors and subspaces can be measured through operators, a *geometry of context* is illustrated in this paper. Once some variables and dimensions of context are selected from the domain for which a context-aware IR tool is designed, the methodology presented in this paper can summarized as follows:

1. for each dimension of context a set of orthogonal vectors is defined — each orthogonal vector of such a set models one factor of the dimension of context;

2. a basis is built for representing a context by selecting one or more factors from each dimension — in this way, a context is modeled by a set of possible contextual factors and one factor refers to one dimension;

3. an informative object is matched against a context by computing a function of the distance between the vector and the subspace spanned by the basis — the closer the vector to the subspace, the more the object is "in the context".

To support the description of the framework let us begin with an example. In Figure 1 we show how the framework represents a document seen from two points of view given by two dimensions. There are two sets of axes — one set of rays is spanned by the vectors $E = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, while the other set is spanned by the vectors $U = \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$; for example, $\mathbf{e}_1$ spans $L(\{\mathbf{e}_1\})$, namely, $L(\{\mathbf{e}_1\})$ is the subspace of the vectors which are obtained by multiplying $\mathbf{e}_1$ by a scalar. A set of coordinates describes a dimension of context; for example, the dimension of context spanned by E may be "document genre". A ray depicts a one-dimensional subspace, e.g. $L(\{\mathbf{u}_i\})$. As a subspace (e.g. a ray) is spanned by a vector, e.g. $\mathbf{u}_i$, the vector describes a value (contextual factor) of this dimension — for example, if the dimension is document genre, $\mathbf{u}_1$ may refer to "introductory" while $\mathbf{u}_2$ may refer to "advanced". In this way, a mathematical representation of contextual factors and dimensions is provided. The mathematical properties of vectors and subspaces can be exploited for representing the properties of contextual factors and dimensions.

In general, factors of distinct dimensions are mutually linearly independent.[2] This means that a contextual factor cannot be described by a vector which is derived by linearly combining the vectors of other factors. In particular, the rays, namely, the vectors corresponding to a given dimension of context are mutually orthogonal for signifying that the values taken by the dimension are mutually exclusive. Orthogonality implies that, the inner product between the vectors spanning the rays is null, thus measuring the event that a mutual exclusion relationship exists between the contextual factors represented — for example, "introductory" excludes "advanced" and viceversa. The way orthogonality can be used as a measure of mutual exclusion is introduced in [30].

Figure 1 depicts how many distinct dimensions co-exist in the same space. This superposition of dimensions can naturally be represented by the infinite sets of coordinates which can be defined in the vector space. In the figure, E superposes U since E and U can at the same time "generate" the vector $\mathbf{x}$ in the same space. The myriad of dimensions model a document or a query from different point of view and each perspective corresponds to a dimension of context. Mathematically, a vector $\mathbf{x}$ is generated by the contextual factors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ as $\mathbf{x} = p_1\mathbf{u}_1 + p_2\mathbf{u}_2 + p_3\mathbf{u}_3$ where $|\mathbf{u}_i| = 1$, $\mathbf{u}_i \perp \mathbf{u}_j$ when $i \neq j$, and $p_1^2 + p_2^2 + p_3^2 = 1$. At the same time, $\mathbf{x} = q_1\mathbf{e}_1 + q_2\mathbf{e}_2 + q_2\mathbf{e}_3$ where $|\mathbf{e}_i| = 1$, $\mathbf{e}_i \perp \mathbf{e}_j$ when $i \neq j$ and $q_1^2 + q_2^2 + q_3^2 = 1$.[3]

Let us consider a set of vectors $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$ where $\mathbf{b}_i$ represents a contextual factor of a dimension of context — as the $\mathbf{b}_i$'s can be of different dimensions, they are independent and not necessarily mutually orthogonal. One projector can be computed from each vector. A projector is an operator that maps a vector to another vector which belongs to a given subspace. A projector is a symmetric and idempotent operator, that is, $\mathbf{B}_i^\top = \mathbf{B}_i$ and $\mathbf{B}_i^2 = \mathbf{B}_i$ — the projectors onto the subspaces $L(\{\mathbf{b}_i\})$'s are defined as $\mathbf{b}_i \cdot \mathbf{b}_i^\top$. If $L(\{\mathbf{b}_i\})$ is the ray containing $\mathbf{b}_i$, then the projection of $\mathbf{y}$ onto $L(\{\mathbf{b}_i\})$ is $\mathbf{B}_i \cdot \mathbf{y}$. If $\mathbf{b}_i$ and $\mathbf{b}_j$ refer to the same dimension, $\mathbf{B}_i \cdot \mathbf{B}_j = \mathbf{0}$ when $i \neq j$ thus defining the notion of projector orthogonality. In general, two projectors $\mathbf{B}_i$ and $\mathbf{B}_j$ are oblique and non-commutative, that is, $\mathbf{B}_i \cdot \mathbf{B}_j \neq \mathbf{0}$ and $\mathbf{B}_i \cdot \mathbf{B}_j \neq \mathbf{B}_j \cdot \mathbf{B}_i$. As there is a one-to-one correspondence between a subspace spanned by a set of vectors and its projector, a projector can be taken as the algebraic operator for a contextual factor and a linear combination of projectors is a mathematical operator which refers to a mixture of contextual factors.

Indeed, a contextual factor is an atomic notion that cannot be decomposed into simpler notions. However, more complex notions can be built by combining contextual factors. Mathematically, the most natural combination which can represent a context is the linear combination. Thus, the operator adopted in this paper is a linear function of projectors by using a predefined set of coefficients which measure the weight of each dimension of context. Therefore, the operator is

$$\mathbf{C}_B = w_1\mathbf{B}_1 + \cdots + w_k\mathbf{B}_k \tag{1}$$

[2]A set of vectors are mutually linearly independent if no vector is a linear combination of the others.

[3]An explanation of these expressions is given in [33].

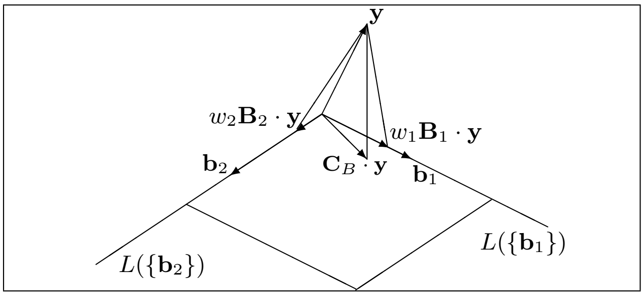

Figure 2: A geometric representation of the ranking function.

where the w_i's are non-negative coefficients such that $w_1 + \cdots + w_k = 1$ and the $\mathbf{B}_i$'s are the projectors onto the subspaces $L(\{\mathbf{b}_i\})$'s. $\mathbf{C}_B$ is called *context matrix* or *context operator* in this paper, since it describes the context described by B. The contextual factors do not need to be mutually orthogonal and thus they can refer to different dimensions.

If the objects are described by the $\mathbf{y}$'s, ranking in context reorders the vectors by the averaged distance between them and the subspaces $L(\{\mathbf{b}_i\})$'s which describe the contextual factors. Therefore, if $\mathbf{y}$ and $\mathbf{C}_B$ are the object vector and the context operator, the ranking function is

$$\mathbf{y}^\top \cdot \mathbf{C}_B \cdot \mathbf{y} \ .$$

From Equation 1, the function becomes

$$\mathbf{y}^\top \cdot \mathbf{C}_B \cdot \mathbf{y} = w_1\mathbf{y}^\top \cdot \mathbf{B}_1 \cdot \mathbf{y} + \cdots + w_k\mathbf{y}^\top \cdot \mathbf{B}_k \cdot \mathbf{y}.$$

As $\mathbf{B}_i = \mathbf{b}_i \cdot \mathbf{b}_i^\top$, it follows that $\mathbf{y}^\top \cdot \mathbf{B}_i \cdot \mathbf{y} = (\mathbf{b}_i^\top \cdot \mathbf{y})^2$, and therefore

$$\mathbf{y}^\top \cdot \mathbf{C}_B \cdot \mathbf{y} = w_1(\mathbf{y}^\top \cdot \mathbf{b}_1)^2 + \cdots + w_k(\mathbf{y}^\top \cdot \mathbf{b}_k)^2 \tag{2}$$

A pictorial description of the ranking function is illustrated in Figure 2. The vector $\mathbf{B}_1 \cdot \mathbf{y}$ is the projection of $\mathbf{y}$ to the subspace (i.e. ray) spanned by $\mathbf{b}_1$ and is scaled by w_1. If $w_1\mathbf{B}_1 \cdot \mathbf{y}$ is summed to $w_2\mathbf{B}_2 \cdot \mathbf{y}$, one obtains a vector which belongs to $L(B)$, namely, the two-dimension subspace spanned by B — this vector is expressed by $\mathbf{C}_B \cdot \mathbf{y}$. Eq. 2 illustrates the degree to which the object represented by $\mathbf{y}$ is close to the contextual factors of B is a weighted average of the size of the projections of $\mathbf{y}$ to the $L(\{\mathbf{b}_i\})$'s.

4. IRF ALGORITHM

In the previous section, a theoretical framework for IR in context has been presented. The framework is generic and can be applied to a range of IR problems. We elect to apply it to IRF because there is an opportunity to capitalize on the framework's ability to handle multiple aspects of interaction (as are visible in interaction context). In this section we describe how the framework can be used to implement an IRF algorithm that captures these aspects. With regard to the experiments described in the next section, we explain (1) how the vectors which represent the contextual factors have been computed, and (2) the specific ranking function used for ranking documents.

The vectors which represent the contextual factors have been computed by Singular Value Decomposition (SVD) of the correlation matrix between the features observed from a set of documents seen by the user during the course of his search. As an example, suppose the following six feature (column) vectors have been observed after seeing six (row) documents:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 3 & 7 & 6 & 7 \\ 2 & 0 & 9 & 7 & 5 & 6 \\ 2 & 0 & 7 & 6 & 4 & 5 \\ 3 & 4 & 8 & 6 & 7 & 7 \\ 4 & 1 & 3 & 6 & 5 & 5 \\ 1 & 28 & 7 & 7 & 5 & 4 \end{bmatrix}$$

where the columns corresponds to, say, (1) display time, (2) scrolling, (3) saving, (4) bookmarking, (5) access frequency and (6) webpage depth,[4] respectively — all of these values may refer, for example, to time or frequencies, and can be seen as of features of user behavior, which is considered as a dimension of context.[5] The following feature correlation matrix is then computed:

$$\mathbf{S} = \begin{bmatrix} 1.00 & -0.42 & -0.14 & -0.78 & 0.11 & 0.05 \\ -0.42 & 1.00 & 0.19 & 0.38 & -0.05 & -0.62 \\ -0.14 & 0.19 & 1.00 & 0.07 & -0.03 & -0.04 \\ -0.78 & 0.38 & 0.07 & 1.00 & 0.00 & 0.00 \\ 0.11 & -0.05 & -0.03 & 0.00 & 1.00 & 0.75 \\ 0.05 & -0.62 & -0.04 & 0.00 & 0.75 & 1.00 \end{bmatrix}$$

An element $s_{i,j}$ of $\mathbf{S}$ is the correlation between columns i and j of $\mathbf{A}$; as a column refers to a feature, the element i, j is the correlation between these two features.[6] To represent the contextual factors we used SVD to compute the eigenvectors of the correlation matrix. The values of an eigenvector are scalars between -1 and $+1$; the further a value is from 0 the more "important" it is. In this circumstance, important means that the feature to which the value corresponds is a significant descriptor of the contextual factor represented by the eigenvector. The value can be likened to an index term weight. As the values may be negative, the sign can express the contrast between features and then the presence of subgroups of features in the same contextual factor. For example, the first eigenvector is $\mathbf{b}_1^\top = (-0.479, 0.516, 0.170, 0.436, -0.308, -0.436)$ and tells that saving is little important ($b_{i3} = 0.170$), while the most important features tend to cluster: scrolling and bookmarking tend to be performed together ($b_{i2} = 0.516, b_{i4} = 0.436$) and tend not to be performed when display time, access frequency, and browsing ($b_{i1} = -0.479, b_{i4} = -0.308, b_{i6} = -0.436$) increase.

Let $\mathbf{b}_i$ be one of these eigenvectors and $\mathbf{y}$ be an unseen document. The function of the distance between the document vector and the subspace spanned by the eigenvector is then used as a measure of the distance between the document and the contextual factor. Therefore, $\mathbf{y}^\top \cdot \mathbf{B}_i \cdot \mathbf{y}$ is computed. If the unseen document vector is, say, $\mathbf{y}^\top = (0.71, 0, 0, 0, 0.71, 0)$, then the distance is 0.31.

The first eigenvector extracted through SVD explains the largest fraction of the variance of the points around their mean vector. It is therefore an average vector interpolating a set of the points corresponding to the seen documents. The eigenvalue is a measure of the variance explained, to be precise, the fraction of variance explained by $\mathbf{b}_1$ is $\lambda_1 / \sum_{i=1}^k \lambda_i$ where $\lambda_i \in \mathbb{R}$ is the i-th eigenvalue.

It is worth noting that $\mathbf{S}$ is symmetric and therefore can be expressed as

$$\mathbf{S} = \lambda_1 \mathbf{B}_1 + \cdots + \lambda_k \mathbf{B}_k \ . \tag{3}$$

This expression means that the relationships between the features are function of the contextual factors thus revealing that the IRF algorithm can discover more information than encapsulated by an average feature vector. The correlation matrices are usually small because the behavioral features do not need to be numerous. Therefore, the computational cost of SVD is quite limited.

In the next section we describe an experiment comparing the IRF algorithm generated using our contextual framework with comparator algorithms that use different types of RF.

5. EXPERIMENTS

The aim of the experiment was to compare the retrieval effectiveness of multiple IRF algorithms that used different sources of implicit feedback and translate this feedback into document rankings. We preferred to measure retrieval effectiveness rather than approximation of assigned relevance judgments (as measured in earlier work [13]) since we were focused on being able to translate our findings directly into end-user utility.

5.1 Methodology

To evaluate the performance of our framework we employed a methodology similar to [11]. The interaction logs of real subjects were used to simulate a user who accesses a series of documents (Web pages) and performs some actions such as reading, scrolling, bookmarking, and saving. The IRF algorithms under investigation are assumed to be part of a system that monitors subject behavior and uses these interaction data as a source of IRF to retrieve and order the unseen documents. When the task or the subject are known, the system records the data by subject / task and then retrieves and ranks the unseen documents for the given subject / task. Although seemingly similar, "topic" and "task" are two different notions. Using Kelly's [37] definitions: "Task was defined for this study as the goal of information-seeking behavior, and topic was defined as the specific subject within a task."[7]

The details of the simulation are as follows:

1. The features of all the documents seen by the user when performing a task and searching for information relevant to a topic are observed. n documents from these are used for computing a representation of context — note that the documents are not ranked by topic at this stage.

2. The observed features of the n documents are used for computing the contextual factors as follows:

   (a) the feature correlation matrix is computed.

[4] The depth of a webpage is the number of links from the root of the website to the webpage itself.

[5] This example is inspired by the dataset used for the experiments reported in this paper.

[6] In general, the number of features is different from the number of documents and then $\mathbf{A}$ will not be square, but $\mathbf{S}$ will always be.

[7] "An example task might be writing a research paper. The topic of this task might be information retrieval and/or interfaces. Another example task might be travel. The topic might be Oregon or Paris. Another task might be shopping, with the topic being shoes or clothes." [37]

(b) the eigenvectors $\mathbf{b}_1, \ldots, \mathbf{b}_k$ are extracted from the correlation matrix — an eigenvector represents a contextual factor.

3. The whole document collection is ranked by the function defined in Section 3 for each projector $\mathbf{B}_i$ at a time. In the experiments reported in the next section no mixture has been investigated and therefore $\mathbf{C}_B = w_i \mathbf{B}_i$ where $w_i = 1$ and $w_j = 0$ for any $i \neq j$. Then, for each projector:

   (a) The ten most frequent keywords of the n top-ranked documents are used for expanding the textual description of the topic, which is then considered as a new, expanded query.

   (b) The expanded query retrieves a list of documents.[8]

   (c) The usefulness scores assigned to the documents are used as ground truth information for evaluating this query expansion-based retrieval.

Normalized Discounted Cumulative Gain (NDCG) was devised as a measure of retrieval effectiveness [39] that was able to handle usefulness scores ranging in a non-binary scale.[9] In this way, NDCG could be a performance metric which is able to make better use of multi-level judgments than precision, which generally must use binary relevance values. In addition to the projector-based method (PRJ) described in Section 4, two other algorithms were also tested:

QRY The topic description was expanded using the query expansion capabilities of MySQL[10] — in this way, a traditional ad-hoc retrieval system equipped with pseudo-relevance feedback (PRF) has been simulated.

CTR The computation of the projectors is replaced with the computation of the unique centroid vector of the cluster of n vectors of the documents seen by the subject when performing a task and searching for information relevant to a topic. That centroid vector has then been used for selecting the feedback documents — the inner product between the centroid vector and the unseen document vectors is then computed for ranking the unseen documents. Note that no clustering is performed.

QRY was chosen as the baseline since (1) it is common practice to compare new RF algorithms to PRF, which is one of the most successful RF techniques, and (2) IRF is a viable substitute for PRF in operational environments, so it is prudent to get a sense of their comparative performance. CTR was chosen because it exploits the same data used by PRJ but aggregates all interaction feature vectors into a single factor, allowing us to determine the value of utilizing multiple factors.

[8]The MySQL full-text functions have been used for these indexing and retrieval tasks.

[9]The discount factor was 2.

[10]It should be recalled that MySQL implements the Vector-Space Model and a version of the weighting scheme described in [40].

5.2 Document Features

The dataset used in this experiment was gathered during a longitudinal user study reported in [37].[11] The set collects the data observed from seven subjects over fourteen weeks and has information about the tasks performed by the subjects, the topics for which the subjects searched the collection, and the actions performed by the subject when interacting with the system. The dataset consists of a set of tuples that each refers to the access performed by a subject when visiting a webpage. The full-text of the documents referred to by the tuples in the dataset has been indexed. The following document features of the dataset were used in our study:

- the unique identifier of the subject who performed the access;
- the unique identifier of the attempted task, as identified by the subject;
- the display time, that is, the length of time that a document was displayed in the subject's active web browser window (display);
- a binary variable indicating whether the subject has added a bookmark for the webpage to the bookmark list of the browser (bookm);
- a binary variable indicating whether the subject has saved a local, complete copy of the webpage on disk (save);
- the frequency of access, namely, the number of times a subject expected to conduct on-line information-seeking activities related to the task (accessfr);
- the number of keystrokes for scrolling a webpage (scroll);
- the depth of the webpage, that is, the number of slashes in the URL (slashes).[12]

In addition to these features, we also have usefulness scores assigned to each document based on how useful it was for a given task for each subject. These scores were assigned by participants in the study based on their own assessment on the usefulness of the document for the task. This gave us document-task-topic judgment tuples that could be used in the assessment of system performance. Some behavioral data were not used because they were elicited from subjects through interview and they could not conceivably incorporated into our search system (e.g., persistence and endurance), are too generic (e.g., task group and week), or have many null values (e.g., stage in task); for details about the data used in this study see [37].

In the next section we present the findings of our study.

[11]2,578 documents were used in the experiments. The dataset contained 2,741 visited documents. The missing 163 were not available for technical reasons.

[12]The number of slashes has been used because it is a measure of webpage quality and is an endorsement of the webpage when the end user selects it. The number of slashes is also known as URL depth and is used for successfully retrieving entry webpages, which are often preferred by the users when finding resources [38].

6. RESULTS

The experimental aim was to test if using a combination of features of user behavior is a more effective means for IRF than established RF models of a centroid of feature vectors. The first question to be answered is whether the model proposed in Section 4 has the potential to retrieve more useful documents than a traditional RF technique — in our investigation, this technique is pseudo-relevance feedback. A first, preliminary answer to this question may be given by Figure 3 which depicts NDCG across all subjects and all tasks for variations in n (i.e. the number of visited documents) and m (i.e. the number of ranked documents used for computing NDCG) for: PRJ, the projector-based method described in Section 3, QRY, the pseudo-relevance feedback *baseline* method, and CTR, the centroid-based method. The NDCG's computed for PRJ were averaged over all the projectors. The values of n and m are distinct — n is used for tuning IRF, m is a parameter for computing NDCG.

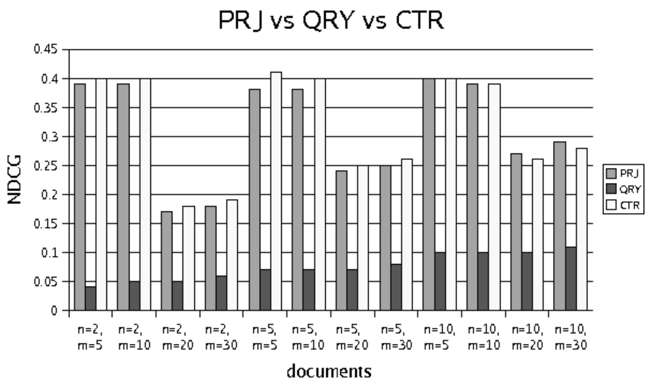

Figure 3: Averaged NDCG for each method and n, m pair.

Figure 3 shows that PRJ and CTR are on average comparable with each other and that QRY is much less effective than PRJ and CTR. The low values for QRY are caused by the high number of ranked document lists without any useful documents among among the top m search results. [13] Figure 3 shows that PRJ and CTR have comparable effectiveness as confirmed by the paired t-tests performed between the NDCG values for the figure; the p-values ranged from 0.44 and 0.91 when computed for the range of n, m values. This may be explained by the fact that CTR uses the same documents as that used by PRJ, yet it is based on the centroid of vectors which is actually an average vector and does not distinguish among the diverse factors by which context may impact on interaction and then on retrieval effectiveness. On the other hand, the potential of PRJ is that the correlation matrix is a linear combination of different projectors which represent implicit contextual factors (Equation 3). As these projectors can be extracted, say, by SVD, PRJ can be refined by varying the projector used for matching documents.

[13]The performance of QRY tends to increase when n increases because of the weak, yet positive correlation between n and the number of useful documents. Indeed, for comparability reasons, the baseline runs were performed when the number of documents per subject/task/topic was higher than $\max\{n, 3\}$.

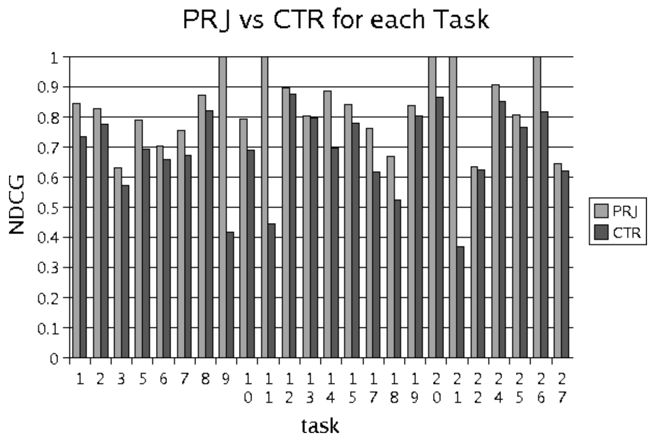

Figure 4: Average NDCG of CTR and maximum average NDCG of PRJ when $n = 2, m = 10$, computed across all subjects and topics, for each task.

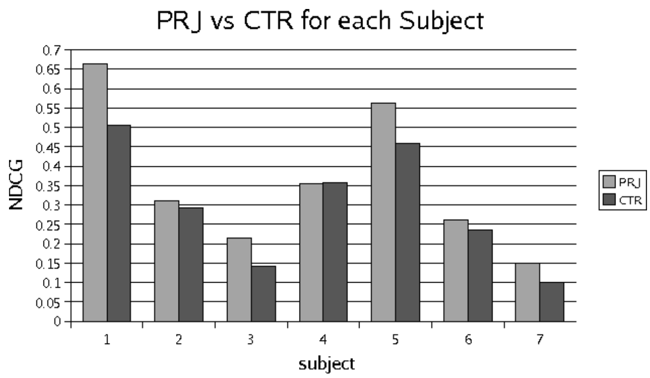

Figure 5: Average NDCG of CTR and maximum average NDCG of PRJ when $n = 2, m = 10$, computed across all tasks and topics, for each subject.

In order to establish the role played by the projectors, an analysis was conducted to compare the effectiveness of CTR with the effectiveness of PRJ by varying the projector. That is, one projector was fixed at a time and the documents were ranked using this projector. We did this for each search task and each subject. Figure 4 depicts the average NDCG of CTR and the average NDCG of PRJ. Each projector produces a different ranking and as a consequence a different NDCG. As our interest was to test whether there exists a projector which outperforms CTR for each subject and for task, the projector which achieved the highest average NDCG of PRJ was selected over all the projectors; therefore, the bars of Figure 4 of PRJ represent the average NDCG of the best performing projector (BPP). The notion of "best" refers to the projector that produces the ranking with the highest NDCG for a subject / task. The value $n = 2$ has been chosen because it is small enough for evaluating the capability of the simulated system to perform effectively even if the feedback is limited; $m = 10$ has been chosen because it is standard practice in IR evaluation to assess retrieval performance up until the 10th retrieved document.

Figure 5 reports on the same result for subjects as reported in Figure 4 for each task. That is, the average ND-

CGs for CTR and the BPP of PRJ are reported for each subject. Figures 4 and 5 suggest that for each subject or task a projector which is more effective than the centroid exists thus indicating that PRJ has the potential of being more effective than a cluster-based method. When PRJ performed better than CTR, e.g., for User 1 or Tasks 9, 11 and 21, the highest weights of BPP correspond to the features of the documents which provided the most effective query expansion terms. This result suggests that a relationship between the BPP and query expansion terms exists. The relationship between BPP and these terms requires further investigation because as it may by symptomatic of a complex interaction between PRJ and pseudo-relevance feedback.

So far we have only considered aggregate performance across all subjects and tasks. However, as Figure 6 shows, the BPP is no consistent across tasks. In Figure 6, a graphical representation of the BPP for each task is given; for example, the BPP of task 1 only consists of access frequency of which weight is 1, while the BPP of task 2 consists of two contrasting groups: one group includes display time and the other group includes access frequency and slashes. As more than one topic may have been used for each subject-task pair, the BPP is an average of the BPPs of each topic of the subject-task pair. As the figure suggests, each interaction feature contributes differently for each of the tasks, and as such each task would need to be represented differently in the IRF algorithm to obtain high retrieval effectiveness. In Figure 7 a similar description is provided for each experimental subject.

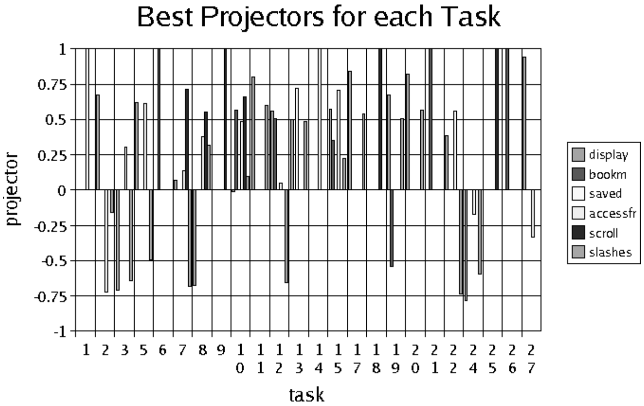

Figure 6: The composition of the BPP when PRJ outperforms CTR as showed in Figure 4.

From the figure it is clear that there are marked differences in the interaction features that comprise the BPP for each subject. This, and earlier presented findings on the BPPs for each search task (in Figure 6) emphasizes the importance of focusing on multiple aspects of user interaction when developing IRF algorithms. In the next section we discuss our experimental results.

7. DISCUSSION

The results indicate that using a combination of features of user behavior is a more effective means for IRF than a traditional pseudo-relevance feedback. Moreover, the results suggest that the selection of contextual factors from interaction may further improve the performance over CTR.

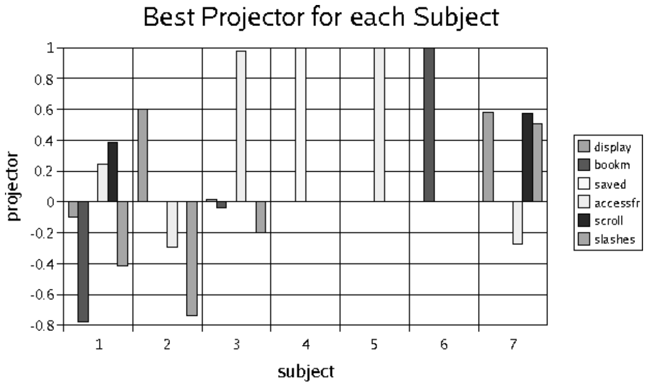

Figure 7: The composition of the BPP when PRJ outperforms CTR as showed in Figure 5.

Figures 6 and 7 also suggest that a combination of features is generally more effective than a single feature such as display time. For example, although the BPP for task n. 1 comprises only accessfr, this is not the case for other tasks. The same is true for subjects. While the bookmarking activity strongly correlates with usefulness for subject n. 6, the BPPs of other subjects involve a combination of features. This outcome makes this paper different from the previous research literature since the correlation between features are not considered as confounding, and they are also regarded as the starting, necessary information for extracting a representation of the contextual factors. This capability is captured by Equation 3.

We have demonstrated that PRJ is more effective, in terms of NDCG, than CTR. This result is somewhat surprising since both use the same volume of information. At first sight, a centroid should be enough for achieving effective IRF and its low computational cost would favor it. When no personalization is required, nor is task adaptability a necessity, CTR is a good solution. PRJ offers the potential to adapt the projector to a user, a task, or both.

The results also suggest that the BPP varies its shape depending on the subject or the task. It was clear that each subject had a unique interaction style when attempting each task, and more than one aspect of this style is necessary to distinguish between subjects. These results suggest that tailoring projectors to users and tasks leads to improved performance over algorithms that do not use such information. This finding is important because justifies the design of IRF algorithms that utilize personalization and task adaptation .

One may of course wonder whether the BPP varies its shape when both subject and task change together. A further analysis of the results has showed that there exists a BPP which is different for each subject-task pair. For example, Figure 8 depicts the shape of the BPP for each subject and task. An important finding of this paper is the existence of an algebraic operator for each subject-task pair which can be used for tailoring document rankings to the user attempting the task. Figure 8 shows another interesting finding: the BPPs for Subjects 3 and 4 are characterized by a single feature; this is consistent for different tasks. This finding can be explained by studying the IRF algorithm. After n documents are visited, some projectors are computed from the feature correlation matrix. Suppose that the BPP is known:

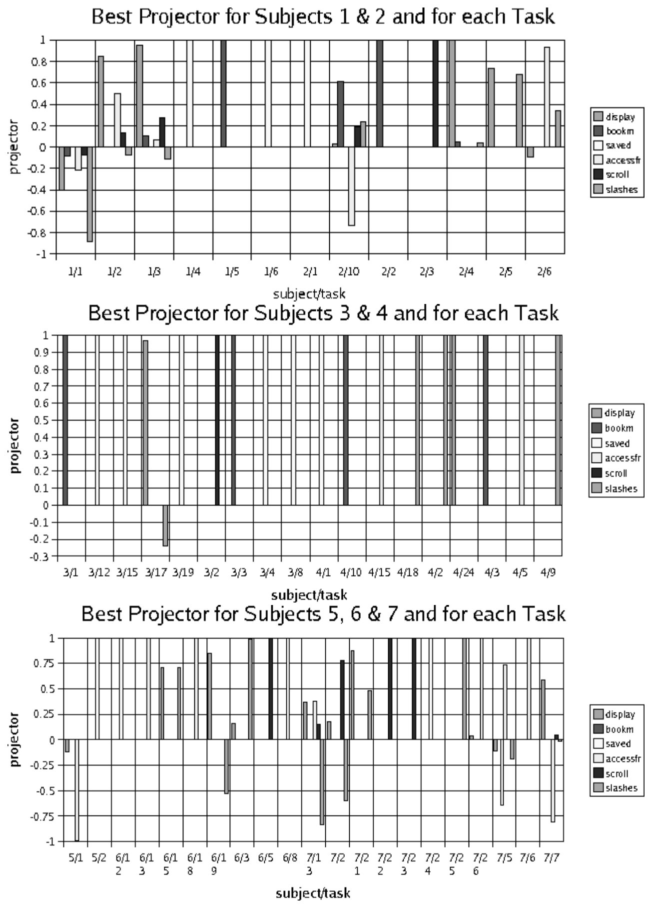

Figure 8: The composition of the BPPs for each subject and task.

when Subjects 3 and 4 are considered, this projector consists of a single feature. When these subjects continue to visit documents, the BPP is able to place the most effective documents at the top ranks, these documents are characterized by the feature of the BPP and contains the most effective terms when added to the expanded query. Table 1 reports the composition of the BPP for each pair subject-task thus making a clear description of the behavior of each subject when attempting a task. The contents of the table demonstrate the diversity of interaction styles evident between users and search tasks even with this relatively small number of human subjects.

8. CONCLUSIONS AND FUTURE WORK

In this paper a new geometric framework that utilizes multiple sources of evidence present in an interaction context (e.g., display time, document retention) has been presented to develop enhanced implicit feedback models personalized for each user and tailored for each search task. These models take the form of projectors or equivalently of eigenvectors extracted from a feature correlation matrix observed from interaction. These implicit feedback models has been compared with alternatives using rich interaction logs (and associated metadata such as relevance judgments) gathered during a longitudinal user study. Two baselines have been used: one based on classical pseudo-relevance feedback where the seen documents were the source for query expansion, the other based on a centroid vector of the seen documents. Our findings demonstrate both the effectiveness of our models and the potential value of incorporating multiple sources of interaction evidence in their development. In particular, it

Subject Task	Eigenvector of the BPP
1/1	bookm (0.08); scroll (0.08); accessfr (0.21); display (0.40); slashes (0.88);
1/2	display (0.85); accessfr (0.50); scroll (0.14); *slashes* (-0.08);
1/3	display (0.95); scroll (0.28); bookm (0.10); accessfr (0.07); *slashes* (-0.11);
1/4	saved (1.00);
1/5	bookm (1.00);
1/6	saved (1.00);
2/1	saved (1.00);
2/2	bookm (1.00);
2/3	scroll (1.00);
2/4	display (1.00); bookm (0.04); slashes (0.04);
2/5	display (0.73); slashes (0.68);
2/6	accessfr (0.93); slashes (0.34); *display* (-0.09);
2/10	bookm (0.61); slashes (0.23); scroll (0.19); display (0.03); *accessfr* (-0.73);
3/1	bookm (1.00);
3/2	scroll (1.00);
3/3	bookm (1.00);
3/4	saved (1.00);
3/8	saved (1.00);
3/12	saved (1.00);
3/15	saved (1.00);
3/17	display (0.97); *slashes* (-0.24);
3/19	saved (1.00);
4/1	saved (1.00);
4/2	slashes (1.00);
4/3	bookm (1.00);
4/5	accessfr (1.00);
4/9	slashes (1.00);
4/10	bookm (1.00);
4/15	accessfr (1.00);
4/18	slashes (1.00);
4/24	display (1.00);
5/1	slashes (0.01); *display* (-0.12); *accessfr* (-0.99);
5/2	saved (1.00);
6/0	bookm (1.00);
6/3	slashes (0.99); display (0.16);
6/5	scroll (1.00);
6/8	saved (1.00);
6/12	saved (1.00);
6/13	accessfr (1.00);
6/15	display (0.71); slashes (0.71);
6/18	accessfr (1.00);
6/19	display (0.85); *slashes* (-0.53);
7/2	scroll (0.78); display (0.18); *slashes* (-0.60);
7/5	accessfr (0.74); *display* (-0.11); *slashes* (-0.19); *saved* (-0.64);
7/6	accessfr (1.00);
7/7	display (0.59); scroll (0.05); *slashes* (-0.01); *accessfr* (-0.81);
7/13	accessfr (0.38); display (0.37); scroll (0.15); *slashes* (-0.84);
7/21	display (0.88); slashes (0.48);
7/22	scroll (1.00);
7/23	scroll (1.00);
7/24	saved (1.00);
7/25	slashes (1.00);
7/26	accessfr (1.00); display (0.04);

Table 1: The composition of the BPPs for each subject and task. Feature subgroups corresponding to negative weighs are italicized. The notation x/y used the first column means "Subject x performed Task y". See Section 4 for an explanation.

was shown that implicit feedback was more effective when the projectors are tailored to the task and personalized to the user. This perspective of multiple information sources of interaction seems to be in line with the ideas of polyrepresentation illustrated in [36].

The following issues are reserved for the future work. Although the IRF algorithms presented in this paper used the documents in the order they appeared in the interaction logs as RF, they did not explicitly consider this order in their computations (as in [41]). Order is important since it could allow the IRF algorithms to incorporate a notion of temporal decay, and give more recent interaction a higher weight when making suggestions. Another interesting extension to the work reported here is an algorithm to automatically select the BPP, i.e., the projector which maximize NDCG for a given pair subject / task. We have shown in this paper that such a projector exists although the selection is as yet an unsolved challenge. A naïve idea would be to select the projector of the first eigenvector. However this is not viable given there is no apparent relationship between NDCG and eigenvalues, and therefore no apparent relationship between NDCG and the variance explained by the eigenvector.

Acknowledgments

The authors are grateful to Diane Kelly for providing the data set used in the experiments reported in this paper.

9. REFERENCES

[1] N.J. Belkin. Helping people find what they don't know. *Comm. of ACM*, 43(8):58–61, 2000.

[2] G.W. Furnas, T.K. Landauer, L.M. Gomez, and S. Dumais. The vocabulary problem in human-system communication. *Comm. of the ACM*, 30(11):964–971, 1987.

[3] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[4] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. *Proc. of SIGIR*, pp. 292–300, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[5] M. Beaulieu. Experiments with interfaces to support query expansion. *J. of IR*, 53(1):8–19, 1997.

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

[7] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. *Proc. of SIGIR*, pp. 272–281, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[8] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. *Proc. of the Int.l Conf. on Intelligent User Interfaces*, pp. 33–40, New York, NY, USA, 2001.

[9] C. Vogt. Passive feedback collection - an attempt to debunk the myth of clickthroughs. *Proc. of TREC*, pp. 141–150, Gaithersburg, MD, 2000.

[10] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. *Proc. of SIGIR*, pp. 19–26, New York, NY, USA, 2006. ACM Press.

[11] R.W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. *Proc.of CIKM*, pp. 297–306, New York, NY, USA, 2006. ACM Press.

[12] J. Teevan, S. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. *Proc. of SIGIR*, pp. 449–456, New York, NY, USA, 2005. ACM Press.

[13] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM TOIS*, 23(2):147–168, 2005.

[14] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. *Proc. of SIGIR*, pp. 154–161, New York, NY, USA, 2005. ACM Press.

[15] D. Kelly and N.J. Belkin. Display time as implicit feedback: understanding task effects. *Proc. of SIGIR*, pp. 377–384, New York, NY, USA, 2004. ACM Press.

[16] R.W. White, I. Ruthven, and J.M. Jose. A study of factors affecting the utility of implicit relevance feedback. *Proc. of SIGIR*, pp. 35–42, New York, NY, USA, 2005. ACM Press.

[17] J. Budzik and K. Hammond. Watson: Anticipating and contextualizing information needs. *Meeting of the ASIS*, pp. 727–740, Medford, NJ, 1999.

[18] R.W. White, I. Ruthven, and J.M. Jose. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. *Proc. of SIGIR*, pp. 57–64, New York, NY, USA, 2002. ACM Press.

[19] J. Kim, D.W. Oard, and K. Romanik. Using implicit feedback for user modeling in internet and intranet searching. Technical report, College of Library and Information Services, University of Maryland at College Park, 2000.

[20] H. Lieberman. Letizia: An agent that assists web browsing. In Chris S. Mellish, editor, *Proc. of IJCAI*, pp. 924–929, Montreal, Quebec, Canada, 1995. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

[21] T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. *Proc. of IJCAI*, pp. 770–777, San Francisco, CA, 1997. Morgan Kaufmann.

[22] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. *Proc. of the Conf. on Uncertainty in Artificial Intelligence*, pp. 256–265, Madison, WI, July 1998.

[23] T. Joachims. Optimizing search engines using clickthrough data. *Proc. of SIGKDD*, pp. 133–142, New York, NY, USA, 2002.

[24] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. *Proc. of SIGKDD*, pp. 239–248, New York, NY, USA, 2005.

[25] R.W. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. *Proc. of SIGIR*, New York, NY, USA, 2007.

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

[27] G. Salton. *Automatic Text Processing.* Addison-Wesley, 1989.

[28] S.K.M. Wong and V.V. Raghavan. Vector space model of information retrieval – a reevaluation. *Proc. of SIGIR*, pp. 167–185, Cambridge, England, 1984.

[29] D. Dubin. The most influential paper Gerard Salton never wrote. *Library Trends*, 52(4):748–764, 2004.

[30] C.J. van Rijsbergen. *The Geometry of Information Retrieval.* Cambridge University Press, UK, 2004.

[31] M. Melucci. Context modeling and discovery using vector space bases. *Proc. of CIKM*, pp. 808–815, Bremen, Germany, November 2005.

[32] M. Melucci. Ranking in context using vector spaces. *Proc. of CIKM*, pp. 477–478, Arlington, VA, USA, November 2006.

[33] M. Melucci. Exploring a mechanics for context-aware information retrieval. *Proc. of the AAAI Spring Symposium on Quantum Interaction*, Stanford, CA, USA, March 2007. AAAI Press.

[34] G. Salton. Mathematics and information retrieval. *J. of Documentation*, 35(1):1–29, 1979.

[35] S. Deerwester, S.T. Dumais, G.W. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. of Am. Soc. Inf. Science*, 41(6):391–407, 1990.

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

[37] D. Kelly. *Understanding Implicit Feedback And Document Preference: A Naturalistic User Study.* PhD thesis, Rutgers, The State University of New Jersey, 2004.

[38] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. *Proc. of SIGIR*, pp. 27–34, Tampere, Finland, 2002.

[39] K. Jarvëlin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM TOIS*, 20(4):422–446, 2002.

[40] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. *Proc. of SIGIR*, pp. 21–29, Zurich, Switzerland, 1996. ACM Press.

[41] I. Campbell. Interactive evaluation of the ostensive model using a new test collection of images with multiple relevance assessments. *J. of IR*, 2(1):85–112, 2000.