Discovering the Preferences of Physicians with Regards to Rank-Ordered Medical Documents

Dympna O'Sullivan[1], Szymon Wilk[2], Wojtek Michalowski[3], Roman Słowiński[2], Roland Thomas[4], and Ken Farion[5,6]

[1] Department of Computer Science, Aston University, Birmingham, UK
d.m.osullivan@aston.ac.uk
[2] Institute of Computing Science, Poznan University of Technology, Poznan, Poland
{szymon.wilk,roman.slowinski}@cs.put.poznan.pl
[3] Telfer School of Management, University of Ottawa, Ottawa, Canada
wojtek@telfer.uottawa.ca
[4] Sprott School of Business, Carleton University, Ottawa, Canada
roland_thomas@carleton.ca
[5] Departments of Pediatrics and Emergency Medicine, University of Ottawa, Ottawa, Canada
[6] Children's Hospital of Eastern Ontario, Ottawa, Canada
farion@cheo.on.ca

Abstract. The practice of evidence-based medicine involves consulting documents from repositories such as Scopus, PubMed, or the Cochrane Library. The most common approach for presenting retrieved documents is in the form of a list, with the assumption that the higher a document is on a list, the more relevant it is. Despite this list-based presentation, it is seldom studied how physicians perceive the importance of the order of documents presented in a list. This paper describes an empirical study that elicited and modeled physicians' preferences with regard to list-based results. Preferences were analyzed using a GRIP method that relies on pairwise comparisons of selected subsets of possible rank-ordered lists composed of 3 documents. The results allow us to draw conclusions regarding physicians' attitudes towards the importance of having documents ranked correctly on a result list, versus the importance of retrieving relevant but misplaced documents. Our findings should help developers of clinical information retrieval applications when deciding how retrieved documents should be presented and how performance of the application should be assessed.

Keywords: Physician preferences, Evidence-Based Medicine, Document Retrieval, Rank-ordered Lists, Information Retrieval.

1 Introduction

As part of our research on clinical decision support systems, we have developed a method for automatically retrieving documents from the Cochrane Library that are relevant in the context of a patient-physician encounter [1]. An evaluation of our method's performance prompted us to reflect on the following question: "*What are*

S. Greco et al. (Eds.): IPMU 2012, Part III, CCIS 299, pp. 142–150, 2012.

physician's expectations and preferences with regards to the rank-ordered presentation of retrieved documents?" Specifically, how do physicians rate the importance of being presented with relevant document on a particular position in a list? Alternatively, how do they value documents that are relevant but misplaced on a list (for example, presented in 2nd instead of 1st place)?

Information retrieval applications that are currently in use return lists of ranked documents where document features are used to estimate a document's relevance for a given query and to compute positions in a ranked list. The established method of evaluating the relevance of documents is to compare retrieved documents with a gold standard for retrieval, which is usually provided by an expert. The effectiveness of the automatic application is then measured in terms of precision – the number of relevant documents a query retrieves divided by the total number of documents retrieved, and recall – the number of relevant documents retrieved divided by the total number of relevant documents that should have been retrieved for the query. However, these metrics do not take into account the position of a document on a rank-ordered list and how physicians perceive mistakes with regard to relevant but misplaced documents. Other measurements such as mean average precision that averages precision over a number of queries has the effect of promoting relevant results closer to the top of a list, however it cannot capture preferences with regard to relevant documents that are out of position on a list. In order to illustrate such an occurrence, assume that for a given query, the gold standard indicates that the correct triple of documents should be [*a*, *b*, *c*], while the information retrieval application retrieved a triple [*b*, *k*, *c*]. Comparing these two triples it can be observed that the retrieval application did not retrieve the most relevant document *a*, it did retrieve a relevant document *b* but placed it in the wrong position (1st instead of 2nd), retrieved an irrelevant document *k*, and retrieved and presented a document *c* in the correct position. All measures of the effectiveness of a retrieval application would focus on the fact that two out of three documents were correctly retrieved while ignoring the order in which they are presented. Such a view would be correct if physicians do not differentiate in terms of the position on which a given document is presented. However, it is not clear if this assumption is correct. Therefore, we studied the following question: *is it correct when evaluating the performance of an information retrieval application to ignore physicians' preferences associated with the order (position) in which documents are presented*? The search for the answer to this question forms the basis of the paper.

The paper is organized as follows. In the next section we briefly discuss research on list-based presentation of search results. In section 3 we describe a study that gathered physician preferences with regard to rank-ordered lists of 3 documents (prior consultations with physicians confirmed that a list with maximum length of 3 documents should be used to present evidence at the point-of-care). Physicians were asked to provide preference information through pairwise comparisons of subsets of rank-ordered lists and these comparisons were analyzed using a GRIP method, which is outlined in the same section. The results of the experiment are presented in Section 4 and the paper concludes with a discussion in Section 5.

2 Background Research

Presenting information as a list is widely used but also widely criticized, because ranked presentation style coupled with the low precision of search engines make it hard for users to find the information they are looking for [2]. In spite of such criticism and subsequent attempts to introduce other methods such as clustering for organizing search results, list-based presentation continues to be the dominant way for organizing information presentation.

Other researchers have studied whether users evaluating list-based presentation follow a depth-first strategy (the user examines each entry in the list in turn starting from the top, and decides immediately whether to open the document in question), or a breadth-first strategy (the user looks ahead at a number of list entries and then revisits the most promising ones) [3]. The results showed that a significant majority (85%), of users relies on a depth-first strategy. Another study used eye tracking (measuring spatially stable gaze during which visual attention was directed to a specific area of the display), to estimate how users process list-based information [4]. The results indicated that users tended to view the first and second-ranked entries right away, and then there is a large gap before viewing the third-ranked entry. A study by Keane et al. [5], also confirmed the inclination of users to access items at the beginning of list. The authors showed that high position on a list often trumpets document's relevance. Considering the potential impact of this inherent user behavior on search results, a school of research is actively devising solutions to overcome the effect of falsely over-promoting web pages by placing them at the top of results list where they will be selected preferentially by users [6, 7].

All these findings have strong implications for the presentation of evidence-based documents to physicians. We hypothesize that if documents presented close to the top of a list have little relevance or are irrelevant, it is likely the entire list will be discarded. While the above statement is confirmed by the research quoted earlier, there is no evidence of how strong physician's preferences are with regards to ordering and positions on rank-ordered lists. Specifically, little is known about how much value they place on receiving relevant documents in the correct order on a list versus how they assess being presented with relevant documents but not necessarily in the right order.

3 Experimental Design

The problem of assessing rank-ordered documents by a physician can be seen as a multiple criteria evaluation problem, where each criterion represents physician's preferences with regards to the relevance of a document presented at a given position on a list. In other words, the value function is a preference model of a specific physician or a group of physicians, which serves to rank a set of rank-ordered lists of documents, taking into account preferences concerning relevance and position of documents on a list. As a preference model, an additive value function can be used, which is a sum of marginal value functions that represent preferences of a physician

on specific criteria.There are many theoretical approaches to estimating an additive value function, including those that rely on the ordinal regression model. In these approaches, preferential information is captured first through the pairwise comparisons of a subset of alternatives (i.e. lists of documents), and then a value function compatible with this information is built [8, 9]. Such a value function represents preferences of a specific user and it can be applied to assess other alternatives that have not been evaluated before.

In our study we used the Generalized Regression with Intensities of Preference (GRIP) method (see [10] for detailed discussion), that derives an additive value function using partial preferential information given by a user in the form of pairwise comparisons of selected alternatives (so-called *reference alternatives*), and ordinal intensities of preference among some of them. It constructs not only the preference relation in the considered set of alternatives, but it also gives information about intensities of preference for pairs of alternatives from this set for a given decision maker. After obtaining results of pairwise comparisons, GRIP checks if any additive value function compatible with the provided preferential information exists. If such a function cannot be found, the method is able to identify pairwise comparisons that prevent representation by an additive value function. Such pairwise comparisons are called *inconsistent* and need to be revised (modified or removed), before proceeding further. Once inconsistencies have been addressed, GRIP constructs marginal value functions for all considered criteria and derives from them an additive value function. This function has to satisfy certain mathematical properties, and because it is computed on a basis of all possible marginal value functions that are consistent with provided preferential information, it is often called a *representative additive value function*. In the analysis presented in the paper we focus only on the marginal value functions associated with the representative function as they provide required insight into physicians' preferences with regards to the retrieved documents.

The experimental design of our study is illustrated in Figure 1. The study consisted of three phases. The first phase started with devising a set of coded triples that represented all feasible combinations of retrieved documents. Each position in a triple, which was considered by GRIP as criterion to be evaluated, was coded as X, N or Y, where X indicates an irrelevant document at a given position, N indicates that a retrieved document is relevant but is placed in an incorrect position on a rank-ordered list, and Y indicates that a relevant document was retrieved and ranked correctly. Thus, for example the triple [*b*, *k*, *c*] mentioned in Section 1 was coded as NXY given [*a*, *b*, *c*] as a gold standard. The coding scheme produced 24 feasible triples out of 27 possible combinations (the smaller number of considered triples is due to some triples being infeasible – i.e. a triple YYN is not feasible because it has two documents that are in the correct position and are relevant, thus the third document cannot be misplaced but can be either irrelevant (X) or correct (Y)).

From the set of 24 triples, a subset of 10 reference triples was selected for 10 pairwise comparisons that corresponded to less obvious evaluations. For example, YYX is intuitively preferred over XYX (retrieving two relevant documents placed correctly on first two positions is preferred over retrieving just one relevant document

that is correctly placed); while comparing NNN with YYX is more difficult (is it preferred that all retrieved documents are relevant but misplaced as opposed to retrieving two documents that are relevant and positioned correctly and a third one that is irrelevant?). Using coded triples for pairwise comparisons allowed us to avoid bias associated with such factors as graphic presentation, or trust in a particular author or publisher.

In the second phase, 6 experienced physicians, all from Ottawa area teaching hospitals, evaluated pairs of reference triples. Study participants represented a range of clinical specialties – emergency medicine, community medicine, internal medicine, intensive care medicine and anesthesiology. All were experienced with using electronic repositories of clinical documents. Prior to the experiment they were informed about the purpose of the study, the experimental design, and how they should conduct pairwise comparisons. Examples of comparisons using triples that were not evaluated in the study were presented and explained. Each physician was asked to independently assess each pair and to state if one triple was preferred over the other, or if they were equally preferred.

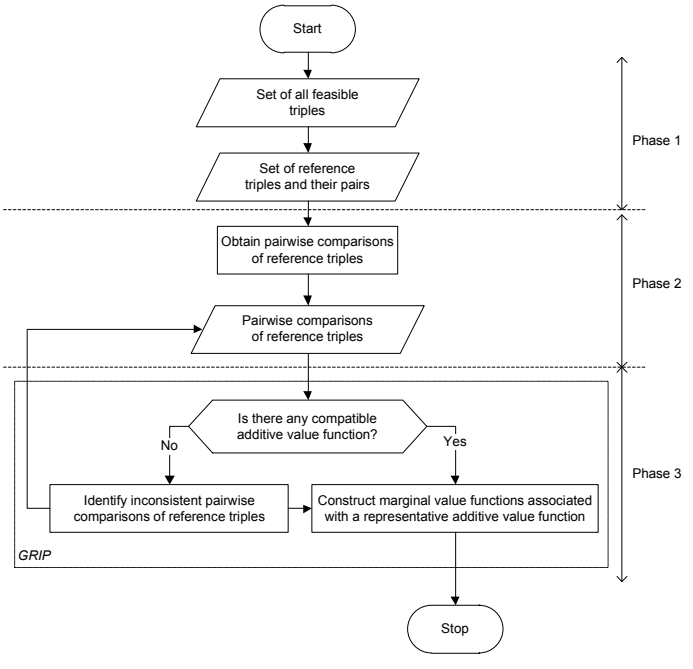

Fig. 1. Experimental design

In the third phase, results of Phase 2 evaluations were used by GRIP to derive three marginal value functions that measure physician preferences regarding the specific position of a document in a triple (1st, 2nd and 3rd).

4 Results

Coded reference triples were presented to the physicians for pairwise evaluation. Table 1 presents the results of the pairwise comparisons of these triples. Each of the physicians (denoted as P1, P2, … P6) was asked to express her/his preferences for one triple (T1) over another (T2) as: T1 preferred over T2 (denoted by symbol "≻"), T2 preferred over T1 (denoted by symbol "≺"), and T1 equally preferred to T2 (denoted by symbol "~"). Responses of physicians P2 and P3 as well as P5 and P6 were identical and therefore are grouped together as (P2_3) and (P5_6), respectively.

Table 1. Physician's pairwise comparisons of triples

T1	T2	P1	P2_3	P 4	P5_6
NNN	YYX	≺	≻	≻	≺
NNX	YXY	≺	≺	≺	≺
NXN	XYY	≺	≻	~	~
NXX	XYX	≺	≺	≺	≺
XNX	XXY	≺	≻	~	≻
XNN	YXX	≺	≻	≺	≺
NNN	YXY	≻	≻	≺	≺
NNX	XYY	≺	≻	~	≻
XNN	XYX	≻	≻	≺	≻
NXX	XXY	≻	≻	≺	≻

The responses presented in Table 1 formed an input for GRIP. The method was applied iteratively to preferential information provided by each physician. First, GRIP identified those responses that no additive value function was able to represent. Such inconsistent responses were removed manually, and then a representative value function able to reconstruct the remaining responses was found. GRIP identified inconsistent pairwise comparisons in responses given by all physicians, except P1; they are marked with grey background in Table 1.

Considering that value domains for each position are discrete (codes X, N and Y), the resulting marginal value function derived as described above becomes a set of the breakpoints. These breakpoints (codes Y and N) are represented in Table 2 for each physician and for each position. The marginal value of code X on any position is equal to 0; therefore it has been excluded from the table.

The analysis of the values in Table 2 provides insights into physician preferences with regards to the presentation of documents. Starting with position 1, all physicians place a high importance on having a relevant document on the 1st position of a list. A correct document (Y) on position 1 receives the highest marginal value across all

Table 2. Breakpoint marginal values at positions 1, 2, and 3 in a triple

Position 1					Position 2					Position 3				
	P1	**P2_3**	**P4**	**P5_6**		**P1**	**P2_3**	**P4**	**P5_6**		**P1**	**P2_3**	**P4**	**P5_6**
N	0.31	0.31	0.1	0.26	**N**	0.19	0.31	0.2	0.26	**N**	0.19	0.19	0.1	0.11
Y	0.42	0.42	0.4	0.53	**Y**	0.35	0.35	0.4	0.32	**Y**	0.23	0.23	0.2	0.16

participants. However physicians are less uniform while placing value on having a misplaced but relevant document (N) on this position. Some are willing to accept misplacement – for example, the analysis of responses provided by P1 and P2_3 for position 1, indicates a smaller drop in marginal values. However, for P4 and P5_6, the difference in marginal values is much more pronounced, indicating that these physicians are less willing to accept on the top position, a document that is relevant but should be ranked lower. Thus, the general conclusion that can be drawn for position 1 is that all physicians want to have the most relevant document in the 1st position on a rank-ordered list.

Moving to positions 2 and 3, physicians are less definitive in their preferences. While all of them value a relevant document (Y) on position 2 higher than on position 3, their preferences are not so definitive with regards to a misplaced document (N) on these two positions. While for P1 there is no difference if a misplaced document is placed on position 2 or 3; however this is not the case for P2_3 and P5_6 who clearly take position into account by assigning higher value to misplaced document (N) if it is on position 2 rather than 3. Responses of P4 fall somewhere in-between – while there is a preference for position 2 over 3, the difference is not that pronounced. In summary, it is possible to conclude that when moving to lower positions, rank order is still important but with diminishing magnitude in the difference between values for correct (Y) and misplaced (N) documents.

The overall conclusion that we draw from the GRIP analysis is that rank order is important for physicians when viewing a list of documents. In particular, it is important that they are presented with a correct document on the 1st position of a rank-ordered list. After position 1, their attitude varies, for some it is still very important that the second most relevant document is correctly placed in position 2, while for others relevance of the documents dominates over ranking for positions 2 and 3 (a document needs to be relevant but can be misplaced). This is coupled with a general reduction in the value of retrieved documents if they are placed on lower positions in a rank-ordered list.

5 Discussion

This paper presented the results of an empirical experiment to model physician preferences with regard to the presentation of rank-ordered list of documents. In particular we wanted to learn if "*it is correct when evaluating the performance of an information retrieval application to ignore physicians' preferences associated with*

the order (position) in which documents are presented? ''Physicians were asked to do pairwise comparisons of rank-ordered triples of documents, and their responses were analyzed using the GRIP method. The results of the experiment show that there is no definitive method of presenting rank-ordered medical documents, however, these general conclusions can be drawn:

- Physicians pay significant attention to the 1st position on a rank-ordered list and they expect that the most relevant document is presented first,
- From a physician's perspective, the importance of presented documents diminishes the lower it is positioned on a rank-ordered list.

These conclusions indicate that the answer to our research question is negative, meaning that it is not correct to ignore order in which documents are presented while evaluating a performance of an information retrieval application.

The obtained results correlate with research on general user searches on the Web (e.g. [4, 5]). They also indicate that when measuring the performance of information retrieval applications it is not sufficient to evaluate only retrieval of correct documents, because physicians clearly put value on the position on a list where a document is presented.

The findings of our study are useful when developing clinical information retrieval applications. They indicate that rank-ordered lists should be short (participating physicians were willing to evaluate lists composed of maximum 3 documents), and that it is imperative to place the most relevant document in the 1st position on a rank-ordered list. However, physicians differ in how they assess subsequent positions – for some having a correctly positioned relevant document in position 2 on a list is very important, while for others, after the first position the relevance of a document gains over its correct positioning.

In future work we intend to use the results of this study in developing a method for more accurately evaluating medical document retrieval. This will translate into revising traditional evaluation metrics such as precision and recall so, for example, a precision value calculated for a document triple YNX will be higher than for a triple NYX (this is not captured when using existing document retrieval performance measures).

Acknowledgment. The authors would like to thank all the physicians who participated in the study. The support of the NSERC-CHRP Program and the Polish National Science Centre (grant no. N N519 441939) are gratefully acknowledged.

References

1. O'Sullivan, D.M., Wilk, S., Michalowski, W.J., Farion, K.J.: Automatic indexing and retrieval of encounter-specific evidence for point-of-care support. Journal of Biomedical Informatics 43(4), 623–631 (2010)

2. Zamir, O., Etzioni, O.: Grouper: A dynamic clustering interface to Web search results. In: 8th International Conference on World Wide Web (WWW 2009), pp. 1361–1374 (1999)
3. Klöckner, K., Wirschum, N., Jameson, A.: Depth- and breadth-first processing of search result lists. In: 22nd SIGCHI Conference on Human Factors in Computing Systems (CHI 2004) (2004)
4. Joachims, T., Granka, L., Pang, B., Hembrooke, H., Gay, G.: Accurately interpreting click-through data as implicit feedback. In: ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), pp. 154–161 (2005)
5. Keane, M.T., O'Brien, M., Smyth, B.: Are people biased in their use of search engines? Communications of ACM 51(2), 49–52 (2008)
6. Cho, J., Roy, S.: Impact of search engines on page popularity. In: 13th International World Wide Web Conference (WWW 2004) (2004)
7. Pandey, S., Roy, S., Olston, C., Cho, J., Chakrabarti, S.: Shuffling a stacked deck: The case for partially randomized ranking of search engine results. In: 31st international Conference on Very Large Data Bases (VLDB 2005) (2005)
8. Greco, S., Slowinski, R., Figueira, J.R., Mousseau, V.: Robust ordinal regression. In: Ehrgott, M., Figueira, J., Greco, S. (eds.) Trends in Multiple Criteria Decision Analysis, ch. 9, pp. 241–283. Springer Science + Business Media Inc., New York (2010)
9. Siskos, Y., Grigoroudis, V., Matsatsinis, N.: UTA methods. In: Figueira, J., Greco, S., Ehrgott, M. (eds.) Multiple Criteria Decision Analysis: State of the Art Surveys, ch. 8, pp. 297–343. Springer Science + Business Media Inc., New York (2005)
10. Figueira, J.R., Greco, S., Slowinski, R.: Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. European Journal of Operational Research 195, 460–486 (2009)