A System for Perspective-Aware Search

M. Atif Qureshi[*†◇], Arjumand Younus[*†◇], Colm O'Riordan[*], Gabriella Pasi[◇], Nasir Touheed[†]
[*]Computational Intelligence Research Group, Information Technology, National University of Ireland, Galway, Ireland
[◇]Information Retrieval Lab, Informatics, Systems and Communication, University of Milan Bicocca, Milan, Italy
[†]Web Science Research Group, Faculty of Computer Science, Institute of Business Administration, Karachi, Pakistan
muhammad.qureshi,arjumand.younus@nuigalway.ie, colm.oriordan@nuigalway.ie, pasi@disco.unimib.it, ntouheed@iba.edu.pk

ABSTRACT

Traditional search engines fail to capture the notion of "perspective" in their search results and at times present the results skewed towards a particular topic. Under most of these cases even query reformulation fails to retrieve desired search results and the underlying reason for such failure is often the bias within the document collection itself (e.g., news articles). A perspective-aware search interface enabling users to look into search results for some "perspective" terms may be of great use for certain information needs. In this paper we describe such a system.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors; H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Human Factors, Performance

Keywords

Perspective, Wikipedia, Bias

1. INTRODUCTION AND RELATED WORK

It is often the case that when using a search engine for information seeking users have an underlying intent [1]. Traditional search interfaces fail to capture the user intent for certain topics and at times return results that may be skewed towards a certain perspective. Here, perspective as defined by the Oxford Dictionary refers to a "point of view"[1] within the search results that may or may not be something what user is looking for. We explain further through the following motivating examples:

- Consider the case of a user who wishes to find more about a certain event (say, a bomb attack in a certain region). The search results returned contain a majority of news reports blaming Islam relating it with terrorism in most of the cases. This prompts the user to explicitly evaluate how much Islam is related to terrorism in the returned search results.
- Consider the case of a user who wishes to find out about roles and rights of women in Islam but the search engine returns articles that contain a high amount of terms highlighting oppression against women instead of women rights and roles. In this case the user is prompted to check the correlation between women and oppression within the search results that have been returned.

Note that the perspective given by most search results (*Islam* in our motivating example (1) and *oppression* in our motivating example (2)) may or may not be aligned with the user's query intent. In case of search results not being aligned with his/her query intent he/she may be interested in observing the amount of perspective tendencies in various news reports.

This paper proposes the concept of a "perspective-aware" search interface that enables the user to explicitly analyse search results for information from a particular perspective with respect to an issued query. To the best of our knowledge, previous research within Human-Computer Interaction and Information Retrieval has failed to capture the notion of "perspective" within the information retrieval process. Early research related to Interactive Information Retrieval by Belkin [2] and Ingwersen [6] suggests the integration of cognitive aspects within the information retrieval process: in line with this suggestion we argue for incorporating the essential cognitive element of "perspectives"[2] within the search engine interface.

Recently the information retrieval community has turned attention to diversification of search results which aims to tackle the issue of query ambiguity on the user side [8]. However, even when formulating a non-ambiguous query users may have an intent that influences the perspective from which the query terms can be interpreted in a text; in case of

[1]This may also be seen as topic drifts within a document.

[2]According to Wikipedia the definition of perspective states the following: "Perspective in theory of cognition is the choice of a context or a reference (or the result of this choice) from which to sense, categorize, measure or codify experience, cohesively forming a coherent belief, typically for comparing with another."

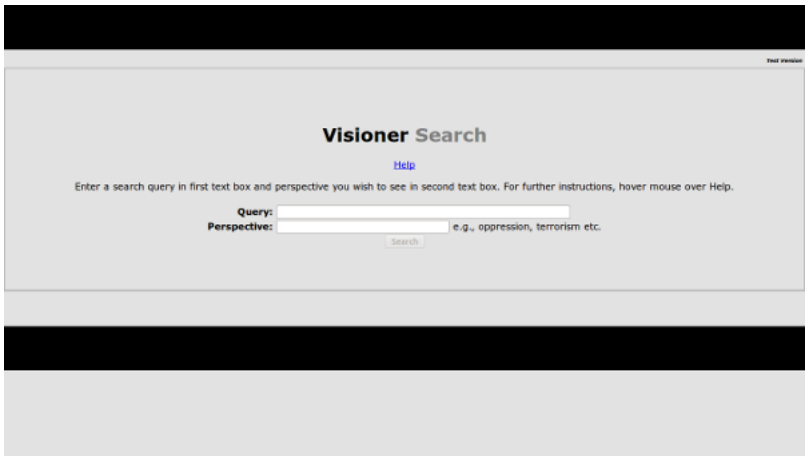

Figure 1: Entry Point of Perspective-Aware Search Interface

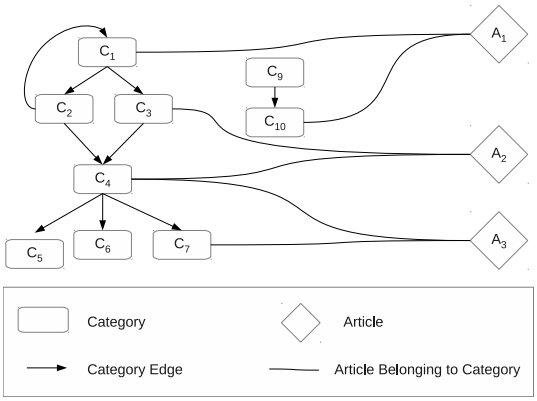

Figure 2: Wikipedia Category Graph Structure along with Wikipedia Articles

perspective mismatch between the user intent and the documents returned in first positions by a search engine, users may find the retrieved results annoying or subjective to a non-agreed perspective [7]. One may argue that a query reformulation technique could be employed to tackle this problem [5]; e.g. considering the motivating example (2), the user could issue a reformulated query such as "roles and rights of women in islam". However, for some topics query reformulation may fail to retrieve the desired search results, and the underlying reason for such failure is often the bias within the document collection itself (e.g., news articles) [10]. Under such a scenario it would be interesting to provide a search interface that would enable the users to look into the search results for some "perspective" terms and we describe such a system in this paper.

2. PERSPECTIVE-AWARE SEARCH INTERFACE AND IMPLEMENTATION DETAILS

This section presents the essential details of the proposed perspective-aware search interface along with the underlying implementation details. We keep the interface as simple as possible on account of research suggesting users' reluctance in switching from a simple search form [3]. Figure 1 shows the entry point of the interface which resembles the standard type-keywords-in-entry-form interface with the augmentation of an additional input text box for entry of perspective terms.

The underlying perspective detection algorithm makes use of the encyclopedic structure in Wikipedia; more specifically the knowledge encoded in Wikipedia's graph structure is utilized for the discovery of various perspectives in documents returned by the search engine. Wikipedia is organized into categories in a taxonomy-like[3] structure (see Figure 2). Each Wikipedia category can have an arbitrary number of subcategories as well as being mentioned inside an arbitrary number of supercategories (e.g., category C_4 in Figure 1 is a subcategory of C_2 and C_3, and a supercategory of C_5, C_6 and C_7.) Furthermore, in Wikipedia each article can belong to an arbitrary number of categories, where each category is a kind of semantic tag for that article [11]. As an example, in Figure 2, article A_1 belongs to categories C_1 and C_{10}, article A_2 belongs to categories C_3 and C_4, while article A_3 belongs to categories C_4 and C_7. It can be seen that the articles and the Wikipedia Category Graph are interlinked and our system makes use of these interlinks for the detection of a certain perspective within a document retrieved by the search engine.

2.1 Underlying Algorithm

The underlying perspective detection algorithm within our system requires the perspective term/phrase to match the title of a Wikipedia article. This may seem to impose a cognitive load on the user at search time. However, this is not the case: as shown in Figure 3 the entered text automatically turns green when a certain user-specified perspective term matches the title of a Wikipedia article, and symmetrically the entered text automatically turns red in case of a mismatch.

Once the perspective term is entered correctly the system fetches the Wikipedia article corresponding to the perspective term referred to as *Seed Perspective Article (PA_{seed})* along with the categories to which it belongs and we use

[3]We say taxonomy-like because it is not strictly hierarchical due to the presence of cycles in the Wikipedia category graph.

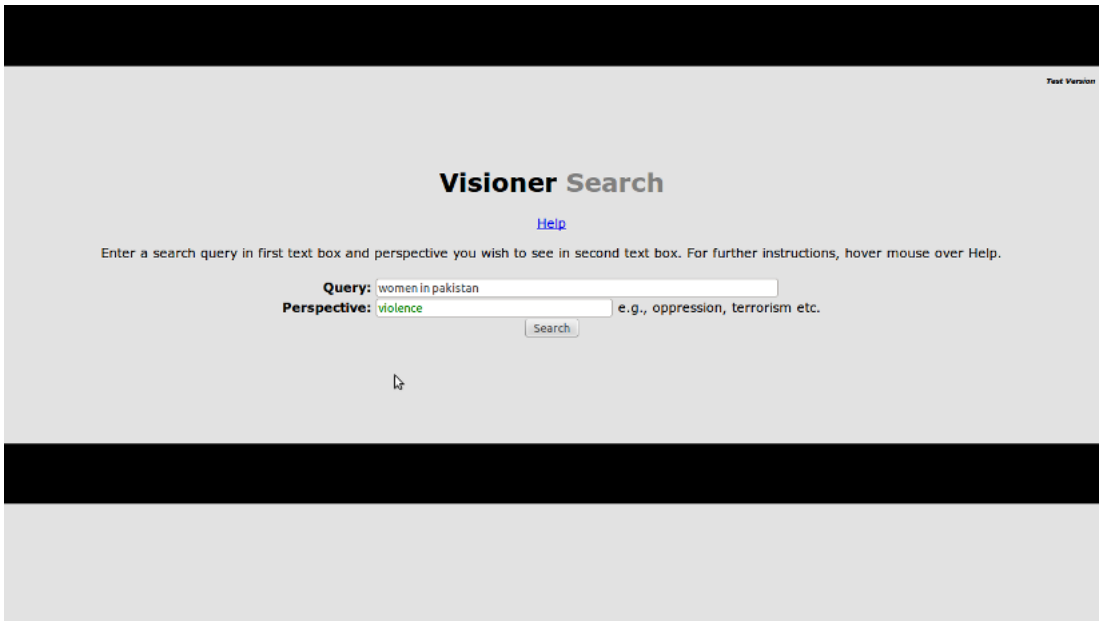

Figure 3: Automatic Text Color Changing to Test Match of Perspective Term with Wikipedia Article Title

PC_0[4] to refer to these categories. After fetching of Wikipedia categories in PC_0, the system retrieves sub-categories of PC_0 until depth 2 i.e., PC_1 and PC_2[5] and collectively these categories related to PA_{seed} are referred to as PC (where PC is union of PC_0, PC_1 and PC_2.). Next, the set of all articles within the Wikipedia category set PC is retrieved and we refer to this set as *Expanded Perspective Article Set* ($PA_{expanded}$). The system then retrieves all categories associated with the set $PA_{expanded}$ which we refer to as WC; note that PC is a subset of WC. Finally, the intersection between PC and WC is retrieved which is a set of categories representative of the domain of the perspective term originally input by the user, we refer to this set of representative categories as RC.

After building the Wikipedia category sets as defined above[6] i.e., PC, RC and WC we match variable-length n-grams within a document with articles in the set $PA_{expanded}$, and we check for cardinality of RC and WC. The cardinality scores along with n-gram frequencies are used to compute a perspective score for each document.

2.2 Search Results Presentation

The perspective scores computed in section 2.1 are displayed within the search results, and based on the perspective score a document receives , we define four levels of perspective adherence as follows: a) High, b) Medium, c) Low, and d) Neutral. Moreover, in case of documents with high, medium and low scores we also report the top-scoring perspective terms that were extracted using the Wikipedia graph structure as explained previously. A sample search corresponding to search query "india pakistan relations" and "terrorism" is shown in Figure 4. As evident from the top search result, there is a high perspective of terrorism within the returned document and perspective terms that our algorithm fetches are as follows: a) the war on terrorism, b) ayman al zawahiri, and c) osama bin laden.

[4]These are basically perspective categories at depth zero.
[5]These are basically perspective categories at depth one and two.
[6]The set building phase is performed through a custom Wikipedia API that has pre-indexed Wikipedia data and hence, it is computationally fast. For details `http://www3.it.nuigalway.ie/cirg/prj/WikiMadeEasy.html`

3. DISCUSSION

There have been many efforts in the information retrieval research to present to users information regarding the relationship between the query and the answer set and the query and document collection. Capturing this information during the retrieval process provides the user with much valuable information (e.g. whether a term is overly specific, or whether a term is ambiguous etc.). Various attempts have been made to tackle this problem, ranging from the definition of snippets to the definition of approaches to cluster search results (Clusty.com), to the presentation of diversified search results in the first position of the ranked list offered to the users. Recently there has been a resurgence of interest in defining visualization techniques of search results that offer an effective and more informative alternative to usual and scarcely informative ranked lists. Pioneer visualization systems are represented by Tilebar [4], and Infocyrstal [9], and these attempts have been aimed to provide the user with more information than that provided by the traditional ranked list.

This additional information can help the user in their search task (e.g. allowing them to navigate the collection more easily or providing evidence to allow the user to reformulate their query more efficiently).

Our proposed system, although related in that we also attempt to give the user an insight into the answer set and its relation to the query, differs in a fundamental manner. Our system, we posit, allows the user to gain insight into the answer set and its relation to the query, but moreover, allows to the user to gain an insight into a *perspective* inherent in the answer set. Our system uses an external and collectively created knowledge resource (which is less likely to be biased

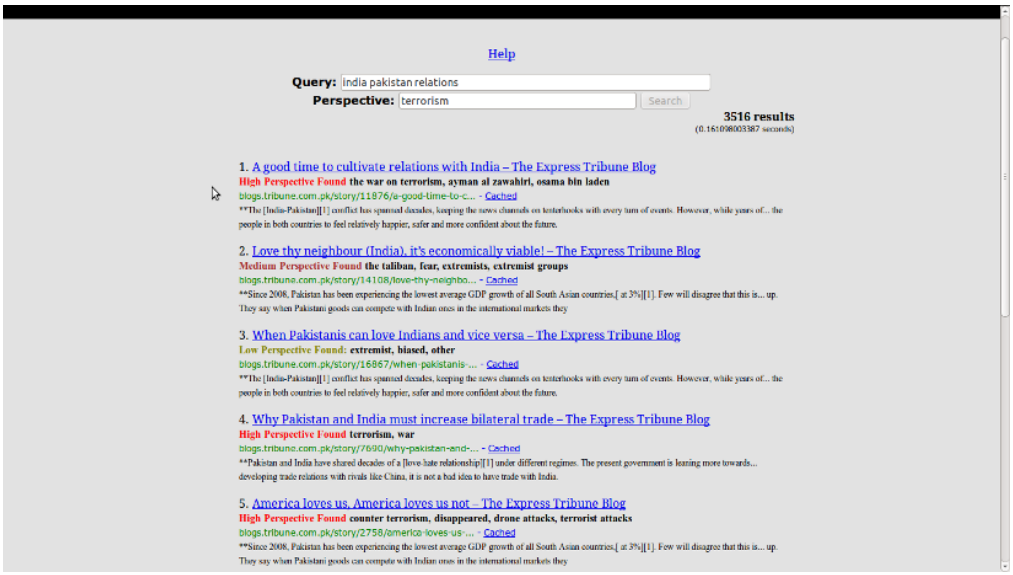

Figure 4: Search Results within Perspective-Aware Search

in a given direction) to obtain extra terms to represent the perspective of interest to the user. This knowledge (perspective term and related terms) does not modify the query (as would an additional query term), but is instead used to highlight the presence of a perspective in the answer set.

In this paper we have proposed a novel approach for capturing the relationship between a user's query and the returned answer set. We do not rely on evidence in the document collection or the query stream, but rather instead extract terms from an external source of evidence to help users quickly see the presence of a particular perspective in the document collection and answer set.

4. FUTURE WORK

Having built the system and undertaken preliminary user evaluations[7], we aim at undertaking a complete and systematic review of the approach. This will comprise a number of separate user evaluation tasks. The initial experiments will involve comparing our search approach with and without the perspective-aware component over a number of tasks to see if the additional context and information provided by our perspective aware system aids the users in a range of information-seeking tasks. Our second planned experiments will be focussed on persons seeking information from newspaper articles, a domain wherein a degree of bias often exists. We wish to explore the users' experience with regards to any perceived bias in the considered corpora.

5. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, 2009.

[2] N. Belkin. Cognitive models and information transfer. *Social Science Information Studies*, 4(2âĂŞ3):111 – 129, 1984.

[3] M. A. Hearst. 'natural' search user interfaces. *Commun. ACM*, 54(11):60–67, Nov. 2011.

[4] M. A. Hearst and J. O. Pedersen. Visualizing information retrieval results: a demonstration of the tilebar interface. In *Conference Companion on Human Factors in Computing Systems*, pages 394–395, 1996.

[5] J. Huang and E. N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 77–86, 2009.

[6] P. Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1):3–50, 1996.

[7] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3):1251–1266, May 2008.

[8] R. L. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 595–604, 2011.

[9] A. Spoerri. Infocrystal: A visual tool for information retrieval & management. In *Proceedings of the second international conference on Information and knowledge management*, pages 11–20, 1993.

[10] A. Younus, M. A. Qureshi, S. K. Kingrani, M. Saeed, N. Touheed, C. O'Riordan, and P. Gabriella. Investigating bias in traditional media through social media. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 643–644, 2012.

[11] T. Zesch and I. Gurevych. Analysis of the Wikipedia Category Graph for NLP Applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, 2007.

[7]The preliminary user evaluations have not been shared in this paper.