Accessibility in Information Retrieval

Leif Azzopardi[1] and Vishwa Vinay[2]

[1] Dept. of Computing Science,
University of Glasgow,
Glasgow UK,
`leif@dcs.gla.ac.uk`,
[2] Microsoft Research Cambridge
Cambridge, UK
`vvinay@microsoft.com`

Abstract. This paper introduces the concept of accessibility from the field of transportation planning and adopts it within the context of Information Retrieval (IR). An analogy is drawn between the fields, which motivates the development of document accessibility measures for IR systems. Considering the accessibility of documents within a collection given an IR System provides a different perspective on the analysis and evaluation of such systems which could be used to inform the design, tuning and management of current and future IR systems.

Keywords: Fairness, Bias, Equality, Accessibility, Retrievability

1 Introduction

Information Retrieval is the area that deals with the storage, organization, management and retrieval of information, where the goal of continual research in the field is to find *better* methods of doing the same. In pursuit of this betterment evaluation has been instrumental in the development of IR Systems. While evaluation has typically focused on the effectiveness [14], or the efficiency [16] of the IR system, these are only two ways in which to assess the quality of an IR system. In this paper, we introduce a complementary view to evaluation which provides a higher level view of the IR system by focusing on the *accessibility* an IR system provides to the documents in a collection.

Accessibility is an abstract concept coined almost 50 years ago in the land use and transportation planning field [10], where it was defined as a measure of potential opportunities for interaction with resources like employment, schooling, shopping, dining, etc. Measuring the accessibility in this context enabled many studies (e.g. [10][12][15][7]) to be performed which examined, for example, how changes in the levels of accessibility to such opportunities affected the urban area (in terms of economic impact, social changes and so forth). The results of such studies provide valuable information to transportation planners and city designers in the development of land use and transportation systems. Before this, planners and designers would focus on measures which were based on

the effectiveness and efficiency of the transportation system (such as, the travel time between particular locations). However, accessibility provided a different perspective, and while related to effectiveness and efficiency, it takes a more abstracted or high level view on the evaluation of transportation systems, considering more general concerns relating to access, instead of focusing on specific instances.

Their definition of accessibility[3] considers the accessibility of opportunities at locations in a physical space (such as a city). The transportation system is the means by which opportunities are made accessible (i.e., the road network and the bus, cycle path and a bicycle, etc). In this context, the main consideration in the design and management of the transportation system is to look beyond efficiency and effectiveness and to consider the accessibility of opportunities given a certain distance or the generalized cost the user is willing to incur to reach these opportunities and the desirability of these opportunities.

In the context of Information Retrieval, an analogy of accessibility can be made as follows. Instead of an actual physical space, in IR, we are predominately concerned with accessing information within a collection of documents (i.e., information space), and instead of a transportation system, we have an Information Access System (i.e., a means by which we can access the information in the collection, like a query mechanism, a browsing mechanism, etc). The accessibility of a document is indicative of the likelihood or opportunity of it being retrieved by the user in this information space given such a mechanism. For example, in a hyper-linked collection exposed by a browsing-based system, a page with no incoming or outgoing links will have no accessibility. Conversely, a page with thousands of incoming links would be very accessible. Here, we consider the accessibility of documents given an IR system, where documents are accessed by querying the system. Each query provides a different ordering in which to access the documents in the information space. Much like a particular bus taking a pre-defined route through a city. However, unlike in the physical space, in the information space, there is no constraint imposed by the user's current location (i.e., at a particular document) because the IR system facilitates access to the collection regardless of location. The IR system is like being at a bus stop where every possible bus route is available, (i.e., the universe of all possible queries), and we can select any route desired, at any time. While this makes every document potentially accessible, the choice of route and distance the user is willing to travel will affect just how accessible documents are in the information space.

---

[3] Accessibility is also a key concept in other areas but defined differently. For instance, the disability rights movement advocates equal access to social, political, and economic life which includes not only physical access but access to the same tools, services, organizations and facilities. Another example is the World Wide Web Consortium (W3C)'s Web Accessibility Initiative (WAI), which is aimed to improve the accessibility of the World Wide Web for people using a wide range of user agent devices, not just standard web browsers. However, accessibility in these contexts concentrates on the physical aspects of accessing the information, and even extends to issues regarding usability and mobility.

In this paper, our main contribution is the introduction of the concept of accessibility and the proposal of how to measure accessibility in this context. To do so, we first describe the related research in Section 2 and draw upon the extensive body of work in transportation planning and land use to provide the basis in developing measures of accessibility for IR system. Then in Section 3, we propose two IR based accessibility measures that are analogous to those in the field of transportation. The introduction of accessibility presents many different possibilities and challenges which can not be fully addressed here, so we summarize this initial contribution in Section 4.

2 Related Work

In Hansen's seminal paper [10] on measuring accessibility in transportation planning and land use, he defines how accessibility could be measured:

> a measurement of the spatial distribution of activities about a point adjusted for the ability and the desire of people or firms to overcome spatial separation. More specifically, the formulation states that the accessibility at point 1 to a particular type of activity at area 2 (say employment) is directly proportional to the size of the activity at area 2 (number of jobs) and inversely proportional to some function of the distance separating point 1 from area 2. The total accessibility to employment at point 1 is the summation of the accessibility to each of the individual areas around point 1. Therefore, as more and more jobs are located nearer to point 1, the accessibility to employment at point 1 will increase.

Key to this definition is the notion that as opportunities become further away the less accessible they become, and that by considering all possibilities to opportunities subject to the cost function based on the distance apart, provides a measure of accessibility. Essentially, this measure quantifies the *potential of opportunities for interaction* [10]. In the context of IR, the opportunities are the documents in the information space, and we wish to capture the *potential of documents for retrieval*.

2.1 Measures of Accessibility in Transportation Planning

There are numerous measures of accessibility that have been proposed in the field of Transportation Planning; the simplest and most popular measures are the Cumulative Opportunity Measures and Gravity Based Measures.

Cumulative Opportunity Measures also known as Isochrone measures count the number of opportunities that can be reached within a given travel time, distance, or generalized cost [15]. An example application of the measure is "the total number of dining opportunities within 400 metres". The advantage of this measure is that it is intuitive and easy to compute. However, the measure is sensitive to the size of the range (around the point of interest) to be considered, and the representation of the opportunities.

First derived by [10], *Gravity Based Measures* provide a general method for measuring accessibility, which is widely used. They differ from cumulative based measures in that they include a cost function within the calculation. Generally, the cost function takes the form of a negative exponential function (as described by [10], above), such that opportunities that are further away will have a lower impact on the final accessibility value. By "further", it is meant in terms of time, distance or generalized cost.

While more sophisticated measures have been developed, such as *Utility Based Measures* [12] and *Activity Based Measures* [7], we shall only be considering the former two methods in this work as they are the most widely used and accepted measures in transportation and planning. Thus, it seems reasonable to use these as a starting point to determine if they can be useful and informative in IR, before developing more sophisticated measures.

2.2   Accessibility in Information Retrieval

Accessibility issues in IR have focused on restrictions (physical and virtual) to index or retrieve information, whether this is because of a physical impairment [8], restricted access due to security clearance [11], or the inability to crawl portions of the web [4]. In each case, documents are inaccessible to the user or the system because of some physical or virtual limitation. For instance, in the latter case, the inability to crawl a web site means that certain documents are not indexed by the IR system, and therefore are not accessible to the user via the IR system. Recently, it was posited that the "searchability" of a web site would be affected by how easily pages can be crawled and how well the search engine matches and ranks them [13]. Searchability and accessibility are therefore very similar concepts. However, we are concerned with the influence of the IR system on accessing documents. Others (e.g. [9][5]) have considered how documents are accessed from the index in the retrieval process to facilitate more efficient retrieval by considering processor, disk and memory constraints. For instance "access-ordered indices" [9] are where the documents which are more likely to be returned at higher ranks are placed before those that are not likely to be returned at higher ranks. Another example, is the caching of queries [3], in web search engines, where results pages are cached in response to popular queries in order to facilitate efficient access.

In essence, IR is all about *accessing information*, and *how the information is accessed*. Our work is focused on measuring the accessibility of documents in the collection given the IR system used to access these documents. This is different from past work, in that we are specifically examining the influence of the IR system to restrict or promote access to the information within the collection as opposed to other restrictions. This paper hopes to establish the idea of accessibility as an integral concept in the field by highlighting its potential in the practical task of developing, building, and optimizing IR Systems, as well as diagnostics and evaluation.

3 Measuring Document Accessibility

Given a collection $\mathbf{D}$, an IR system accepts a user query $\mathbf{q}$ and returns a ranking of documents $\mathbf{R_q}$, which are deemed to be relevant to $\mathbf{q}$ from within $\mathbf{D}$ by the IR system. We can consider the accessibility of a document as a system dependent factor that measures how retrievable it is, with respect to the collection $\mathbf{D}$ and the ranking function used by the IR system. Using the analogy of transportation, entering a query is like to choosing a particular bus, where the order of documents returned are like the order of destinations reached for that given bus route. Opportunities to interact with resources while traveling along the route are reflected by going through the documents returned in the ranking $\mathbf{R_q}$. The accessibility of the resources (i.e., documents) is dependant on the willingness of the user to travel a certain distance along the route (i.e., traverse down the ranked list) and all the queries that users are likely to travel along. So, by adapting the measures from transportation planning, we propose a general measure of the accessibility of a document, as:

$$A(\mathbf{d}) = \sum_{\mathbf{q} \in \mathbf{Q}} o_q \cdot f(c_{dq}, \theta) \tag{1}$$

where o_q denotes the likelihood of expressing query $\mathbf{q}$ from the universe of queries $\mathbf{Q}$ and $f(c_{dq}, \theta)$ is a generalized utility/cost function where c_{dq} is the distance associated with accessing $\mathbf{d}$ through $\mathbf{q}$ which is defined by the rank of the document, and θ is a parameter or set of parameters given the specific type of measure.

A cumulative based measure can then be defined as follows: $\theta = c$, where c denotes the maximum rank that a user is willing to proceed down the ranked list. The function $f(c_{dq}, c)$ returns a value of 1 if $c_{dq} \leq c$ (with the top-most position considered as rank 1), and 0 otherwise. So, if returning a document in response to a given query has a distance greater than c associated with it, then it is considered unaccessible (for this query). For another query however, the document may be accessible because the cost of accessing it is within the distance c. Alternatively, the document could be considered accessible for the same query but to a user who has a higher cost threshold. Since all the documents within the cutoff defined by c are equally weighted, this type of measure emphasizes the number of times the document can be retrieved within that cutoff over the set $\mathbf{Q}$.

A gravity based measure can also be defined by setting the function to reflect the effort of going further down the ranked list, such that the further down the ranking the less accessible a document becomes. There are numerous ways in which such a function could be determined. Here, we adopt the function suggested in [10], where the accessibility of the document is inversely proportional to the rank of the document, such that:

$$f(c_{dq}, \beta) = \frac{1}{(c_{dq})^\beta} \tag{2}$$

where, the set of parameters θ includes β which is a dampening factor that adjusts how accessible the document is in the ranking. Interestingly, if the β parameter is set to one, then accessibility of the document for the given query is equivalent to the reciprocal rank of the document, which is related to the (expected) search length [6]. When there is only one relevant document, the expected search length is equivalent to the reciprocal rank of the document. Intuitively, the expected search length (ESL) and accessibility of documents is related, because the expected search length corresponds to how many irrelevant documents have to be examined in order to find the relevant documents. The expected search length to a particular document is proportional to the accessibility of the document for a given query. However, what the accessibility measure captures is more general, i.e., how retrievable the document is given all possible/likely queries regardless of relevancy, but this link to ESL and reciprocal rank appears to provide a connection between accessibility and effectiveness. As we have previously mentioned this direction is left for future work.

Given either measure, $A(\mathbf{d})$ provides an indication of the opportunity of retrieving $\mathbf{d}$. This value can be obtained for each document $\mathbf{d} \in \mathbf{D}$ so that we can compare whether there is more opportunity to retrieve one document over another. Using this measure to compare groups of documents has potential to aid in the design, management and tuning of retrieval systems in a number of ways. Imagine that for a given collection of documents and a given IR system, the average $A(\mathbf{d})$ of a set of documents is extremely high, while for another set of documents the average $A(\mathbf{d})$ is very low. Perhaps, the first set of documents was a group of site entry pages, and our system has a prior towards such pages, thus we would expect these pages to have a higher $A(\mathbf{d})$. In this case, it is desirable that these documents are so accessible. On the other hand, if the set of highly accessible pages was composed of spam pages, because these pages have used "tricks" to artificially inflate the number of queries for which they are retrieved, then this is not desirable and the system needs to be adjusted. Alternatively, if there is a set of documents which are virtually inaccessible in the collection, then it is a management decision to decide whether these documents should be included in the index or not.

At a higher level, the measure $A(\mathbf{d})$ motivates questions regarding how accessible documents in the collection should be, and whether we are interested in trying to "hide" or "promote" certain documents within the collection. Or whether we should adopt an approach that ensures access to the information is free from bias, i.e. "universal access"[4] so that *any document is as accessible as any other document* in the collection. This provides a novel framework for measuring document accessibility, which enables the consideration of such questions and issues.

---

[4] As previously mentioned, the disability rights movement advocates equal access and terms this notion as universal access.

4 Conclusion and Future Work

The main contribution of this paper is the introduction of the concept of accessibility and quantifying the accessibility of documents in the collection given a particular IR system. Measures of accessibility are not performance measures like effectiveness or efficiency, but instead are measures of the *potential of documents for retrieval* (a.k.a., their retrievability [2]). This abstraction provides a novel way to quantify and detect different levels of accessibility within the collection imposed by the IR system. For a system administrator, this could prove to be very useful in designing, managing and tuning the IR system (see [1] for empirical examples).

This work represents the initial step towards formalizing accessibility and developing accessibility measures for information spaces, in IR and more generally for any Information Access system. However, there are many open problems, challenges and issues which have arisen as a result of this work. Further research needs to be conducted in two main directions:

1. the calibration, computation and estimation of document accessibility measures, and
2. the application of document accessibility measures.

Acknowledgements The authors would like to thank Stephen Robertson, Keith van Rijsbergen and Murat Yakici for their helpful and insightful comments and suggestions.

References

1. L. Azzopardi and V. Vinay. Document accessibility: Evaluating the access afforded to a document by the retrieval system. In *Workshop on Novel Methodologies for Evaluation in Information Retrieval*, pages 52–60, March 2008.
2. L. Azzopardi and V. Vinay. Retrievability: an evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 561–570, 2008.
3. R. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri. The impact of caching on search engines. In *Proceedings of the 30th ACM SIGIR conference*, pages 183–190, 2007.
4. P. Bailey, N. Craswell, and D. Hawking. Chart of darkness: Mapping a large intranet. Technical report, CSIRO Mathematical and Information Sciences, 2000.
5. S. Buttcher and C. L. A. Clarke. A document-centric approach to static index pruning in text retrieval systems. In *Proceedings of the 15th ACM Conference on Information and Knowledge Management*, 2006.
6. W. S. Cooper. Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science*, 19(1):30–41, 1968.
7. X. Dong, M. E. Ben-Akiva, J. L. Bowman, and J. L. Walker. Moving from trip-based to activity-based measures of accessibility. *Transportation research. Part A, Policy and practice*, 40(2):163–180, 2006.

8. I. Fajardo, J. J. Canas, L. Salmeron, and J. Abascal. Improving deaf users' accessibility in hypertext information retrieval: are graphical interfaces useful for them? *Behaviour and Information Technology*, 25(6):455–467(13), 2006.

9. S. Garcia, H. E. Williams, and A. Cannane. Access-ordered indexes. In *Twenty-Seventh Australasian Computer Science Conference (ACSC2004)*, pages 7–14, 2004.

10. W. Hansen. How accessibility shape land use. *Journal of the American Institute of Planners*, 25(2):73–76, 1959.

11. D. Hawking. Challenges in enterprise search. In *ADC '04: Proceedings of the 15th Australasian database conference*, pages 15–24, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

12. H. Neuburger. User benefit in the evaluation of transport and land use plans. *Journal of Transport Economics and Policy*, 5:52–75, 1971.

13. T. Upstill, N. Craswell, and D. Hawking. Buying bestsellers online: A case study in search & searchability. In *7th Australasian Document Computing Symposium*, Sydney, Australia, 2002.

14. C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition edition, 1979.

15. M. Wachs and T. G. Kumagai. Physiscal accessibility as a social indicator. *Socioeconomic Planning Science*, 7:327–456, 1973.

16. I. H. Witten, A. Moffat, and B. T. C. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishing, San Franciso, second edition edition, 1999.