Ranking – Use and Usability

Bettina Berendt

Abstract

Ranking has recently attracted much attention, conceptually as well as algorithmically and in its uses in major Internet search engines. A core goal of ranking and related techniques on the Web is to help people find what they are looking for – and (depending on the application) to suggest to them things that they didn't know they were looking for, but might still find interesting. Any evaluation of these techniques should therefore consider such deployment scenarios. The paper starts from the observation that search engines and recommender systems generally provide users with some ranking on resources, and that standard evaluations rest on a comparison of this system output with an assumed mental representation of the user's "true ranking". This is followed by an overview of i) where and how ranking is used by the operators of a Web site or similar service, ii) how ranking is used by the end users of that site or service, and how such usage is measured, and iii) how and according to which criteria this usage and the success as well as the quality of ranking are measured. The paper demonstrates how an interdisciplinary approach can sharpen the view of challenges and promises of this user-oriented analysis of intelligent information access.

1 Introduction

Ranking has recently attracted much attention – conceptually, see for example (1; 34; 47), as well as algorithmically and in its uses in major Internet search engines, including PageRank (30) and other algorithms (11). In the Belgian context, the 2008 symposium "The mathematics of ranking"[1] assembled a wide range of perspectives on the topic. The present paper is based on an invited talk at that symposium and on the perspective of a Web mining researcher. The aim of both

[1] `http://www.cs.kuleuven.ac.be/~nalag/research/workshops/ranking/`

Received by the editors June 2009.
Communicated by A. Bultheel.
2000 *Mathematics Subject Classification :* 91E45; 68U35; 68T05; 91E10; 68P10.
Key words and phrases : Ranking, rating, evaluation.

the talk and the paper is to open a discussion on the uses of ranking, by giving an overview of how rankings are used today (focusing mostly on the Internet), how they are evaluated with respect to their usefulness for their human users, and which – mathematical and other – challenges arise in this process.

Evaluating a ranking is asking the following question: When is a ranking, returned by a search engine or similar program, "good"? In this paper, I start from the postulate that the ultimate purpose of any such ranking is to help users, and I describe approaches and challenges for evaluating rankings with regard to this top-level purpose. The core question then becomes: Is the ranking the "right ranking" for the given person, in the given circumstances? This leads to further questions such as: how and for what is the ranking used, and is it presented in a usable way?

In the present paper, I will investigate the evaluation of rankings with a view to answering these questions. The techniques will be illustrated with examples from two main classes of applications that order their result sets: search engines and recommender systems. The focus will be on the two main ways in which orderings are currently expressed: rankings and ratings. The paper is intended as an introduction to today's main approaches for evaluation. Its contribution is the identification of eight *challenges*: Starting from a simple formal model of ranking and evaluation, I highlight the plethora of implicit assumptions behind this model. I describe evidence, mainly from behavioural studies, that shows that these assumptions are often not satisfied in real-world usage contexts. The paper outlines formal and other modifications that have been proposed to deal with these challenges, but it also shows that such solution proposals exist only for the first challenges. The second intent of this paper is therefore to inspire further research into formal and less formal solutions to the many challenges of user-oriented evaluation. For reasons of space, the article is designed as exemplary rather than as a comprehensive survey.

The remainder of the paper is organised as follows: Section 2 lays the groundwork in the form of basic terminology and definitions, and it presents the simple model of evaluation that will be made progressively more complex in the remaining sections. In Sections 3 to 10, eight main challenges to the simple model are described with reference to empirical evidence, together with solution approaches. Section 11 concludes with an outlook.

2 The basic setting: A simple model of aligning two rankings

In a first step, a basic framework has to be found to be able to answer the question "Is this (system-generated) ranking the right ranking for a user?"

A simple model for answering this question derives from Information Retrieval (3).[2] It assumes that given a task (specified, for example, by a search

[2]This model is a deliberate simplification, done for the purposes of discussing the challenges with respect to one common reference point. While the simplification does describe evaluation measures and procedures that are often used in studies of ranking and rating methods, the development and use of more sophisticated evaluation methods is an active research area in Information Retrieval, cf. (8) and the articles in that Special Issue.

query), there are a system-generated set of items and a "ground-truth" set of items that are "really" relevant to the task. This "ground truth" is often established by experts. For the sake of simplicity, in this paper I ignore the distinction between external experts and (possibly expert, possibly non-expert) information seekers, and postulate that the ground-truth set of items is a mental structure in the user's head. (Simplifying, one could assume that with the benefit of hindsight, the user would recognise this to be her "true" preferences.)

System preferences and user preferences These two sets of items will be called S (system preferences) and U (user preferences), respectively. The sets may have different structures defined on them. Most common in S are order relations that correspond to $>$ (*rankings*), as for example in Internet search-engine interfaces, or to $\geq$ (*ratings*), as for example in many Internet movie or music recommender systems. The ratings are often numbers from, e.g., 1 to 5 associated with the items; in evaluations, it is usually assumed that these numbers are interval-scaled measures with equal distances between subsequent numbers. The U sets are treated differently by different evaluation measures, generally as non-structured or as structured like the S sets.

In evaluations, one usually starts from concrete data for a given task, i.e. S and U are extensional descriptions of the preferences. To the extent that one has access to the internals of the search engine, one can in addition inspect the system's ranking function s, which is usually a function of the user input (e.g., a search query) and of valuations of multiple criteria, such as the potential results' textual content, metadata, etc., cf. (3).

Conceptually, a function producing a ranking or rating can be described as a two-step process:

In a first step, a function $score(x)$ is computed for all items $x \in X$ to be considered. Usually, $score : X \mapsto \mathbb{R}$. This function relies on the features a_j of x, which are weighted by valuations v_j:

$$score(x) = f(v_1(a_1(x)), v_2(a_2(x)), \ldots v_m(a_m(x))) \tag{1}$$

Valuations may depend on different features simultaneously, and the score of an item in most cases depends also on a user input such as a query or the user ID, as well as on the whole set of item alternatives. In the following, notation is simplified by assuming per-feature valuations and dropping the query and alternative-set parameters. In addition, it is assumed that higher scores mean better results.

In a second step, a decision function $s(x)$ transforms these scores and maps the items into

- the range $\{0, 1\}$ if the aim of s is to filter *relevant items* and S is an unstructured set. A possible definition is: $s(x) = 1$ if $score(x) \geq \theta$ for some threshold $\theta \in \mathbb{R}$, and 0 otherwise.
- a range like $\{1, \ldots, 5\}$ if the aim of s is to produce a *rating*. A possible definition is: $s(x) = r$ if $score(x) \in [\theta_r; \theta_{r+1})$ for thresholds associated with each rating value r.

- a range in $\mathbb{N}$ if the aim of s is to produce a *ranking*. Here, the elements of X are ordered by their scores. Subsequently, in most cases simplifications are applied in order to produce a ranking that is easily understandable for human users, i.e. a list in which items stand in a one-to-one relation with ranks:

  1. All elements with scores below a minimum relevance threshold are disregarded.
  2. The remaining relevant elements are ordered such that ties (equal scores) are broken by some deterministic or random process.
  3. The magnitudes of differences between successors on the list are disregarded.

  As a result, all ranks between 1 and the number of relevant items are filled.

***Ranking* as a general term** For simplicity of notation and where confusion can be avoided, I will use *ranking* to denote any of the structures on U and S.

Obtaining the data S can be obtained by running the given system's algorithms on the required input, generally a search query ("get items relevant to this query") or a user ID ("get items relevant for this person"). As an assumed mental representation, U cannot be obtained directly. Instead, standard methods of elicitation are used, and their results are treated as U. The main methods are to ask people (in interviews, with questionnaires or tasks, ...) and to rely on these self-reports, or to observe their behaviour and to draw inferences towards the preferences from these observations. Standard methods are logfile analysis (choosing an item is interpreted as a sign of preference; differentiations can be made depending on the type of choice, including clicking, downloading, or buying) and eye-tracking. The latter measures overt visual attention, which can be a good indicator of preference (41).

Evaluating the fit: Accuracy measures To evaluate the fit, the sets/lists S and U are compared. The more similar they are, the better the ranking is considered to be.

Classical measures for internally unstructured sets are

$$precision = \frac{\mid S \cap U \mid}{\mid S \mid} \tag{2}$$

$$recall = \frac{\mid S \cap U \mid}{\mid U \mid} \tag{3}$$

and their combinations like

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

To take into account the ranking on S without making assumptions about the internal structure of U, measures like the following are popular:

$$precision@n = \frac{\mid (\text{top } n \text{ items of } S) \cap U \mid}{n} \tag{5}$$

Measures like the mean absolute error or (rank) correlations also consider the structures on S and U:

$$MAE = \frac{\sum_{i=1}^{N} \mid s_i - u_i \mid}{N} \tag{6}$$

$$corr = \frac{\sum(s - \bar{s})(u - \bar{u})}{n * stdev(s) * stdev(u)} \tag{7}$$

where s or s_i is short for $s(x_i)$ and u or u_i is short for $u(x_i)$, $\bar{\bullet}$ and $stdev(\bullet)$ denote the mean and empirical standard deviation, and N is the total number of items compared. *corr* is the Pearson correlation when the s and u are ratings and the rank correlation when they are ranks.

A large number of such measures exists today, for examples see (21; 54). Not all of them have to be investigated if one wants to evaluate the fit of a given S and U, since in general various measures appear to select the same properties of accuracy and fall into a small number of 'equivalence classes', see for example (21; 55). For some measures, equivalence has not only been demonstrated in simulations, but also proved (15). Robertson and Zaragoza (45) have proposed to optimise such performance measures: to specify their desired properties and then learn the best measure from data.

However, a closer look reveals that the focus on goodness-of-fit needs to be enhanced by more detailed considerations of users and usage. Various features of these present *challenges* to the simple model of fit and evaluation.

3 Challenge 1: The user's ranking depends on the context

The problem Imagine a search for "star coiffeurs" that returns a ranked list of hairdressers' stores in Kansas City, San Jose, and other places, all featuring their special styling proposals for women. This may be (a) not very helpful if the searcher is a man; (b) not at all helpful if the searcher is currently in Brussels, looking for somewhere to get his/her hair cut; (c) very helpful if the searcher looks for information for a sociological project on the marketing of high-end hairstyling on the Internet. The three cases illustrate three types of factors that often have a strong influence on user preferences: (a) persistent properties of the user, (b) transient properties of the user, or current properties of the environment of search, and (c) the current goal/task of the user. These three types of factors are jointly known as *context*, cf. for example (13) and the articles in that Special Issue. Some aspects of context are relational. For example, people may have different preferences for information depending on whether it is in their native language or not, irrespective of the identity of that language, cf. (6).

Solution approach: measure context, adapt the ranking function A straightforward solution approach to this challenge is to integrate context features that have been recognised as important and that can be measured into the ranking function.

Thus, formally, the scoring function in Equation (1) needs to be enhanced by terms depending on features a that are context (e.g. user) features and by arguments that describe the relation between item and context features. Simple examples of transient and persistent user features include location and gender.

The measurability of these different factors varies, as does their availability for adapting $s()$. A user's location can be retrieved from the user's IP address at the level of the country and/or region within a country. For example, the producers of the MaxMind database state that their localisation is "99.8% accurate on a country level, 90% accurate on a state level, 83% accurate for the US within a 25 mile radius" (35). Mozilla's Geode browser plug-in uses Wifi signals; it "works both inside and outside with an accuracy of between 10 to 20 meters, normally within a second" (36). Mobile devices can now be localised within ca. 100 metres based on cell information (within cities), ca. 50 metres based on GSM fingerprints, and below 10 metres indoors based on further fingerprinting technology (53). While IP addresses are (nearly) always available, other location signals may or may not be available.

Measuring other features like gender is generally less straightforward; the usual approach is to learn a classifier with data-mining algorithms from prior behavioural data like the clickstream of a user, e.g., (22). The classifier is then applied to a new user's profile and/or behavioural data in order to predict whether this new user is male or female. Other data sources include self-profiling information of registered users. These are available in many rating sites, and also in search engines with customisation options like iGoogle[3].

Investigating native language as a relational feature of context, Kralisch and Berendt (28) showed that native language can be inferred from IPs and geolocation, and in a validation study obtained a small error bound on this technique (6.7%), but more large-scale studies are needed to replicate this finding.

The last question is how to adapt $s()$ to take these factors into account. The assumption of location-based services is that geographically close results are more relevant (this requires also the measurability of the result items' locations, which is usually possible by their address). Another heuristic could be that search on a (small) mobile device is more likely to ask for local information, while search on a PC may also be for research purposes. One example of this is the common search-engine option to search "only for pages from [your local country]", which can but need not be checked. Deriving valuation functions $v(a(x))$ for other features a like gender or task in order to re-rank items x by whether they will be liked by people with the detected gender or task, is usually much less straightforward; the usual approach is to learn these functions from historical data, in which members of a certain class (e.g., male or female) showed evidence of liking/preferring certain content. Then, simple classifiers that, when applied to an item, predict whether it will be liked or not, are learned (32; 22); alternatively,

[3] `http://www.google.com/ig`

ratings or rankings may be predicted.

Further context features belong to the environment rather than to the user. They can often be measured/inferred and used for adapting $s()$ in similar ways. For example, the IP address could be used to derive an estimate of whether the person searches from home or work, or location information could be enriched to assess the local weather, and these estimated values of context variables could be used to better adapt the ranking to the context.

The trade-off between user-adaptivity and privacy The measurement and use of information on the user or her immediate environment may help a system deliver better results for the user in her current context, even though this cannot guaranteed (for example, a man may look up a hairdresser for his wife). However, the measurement and use of personal data may constitute a significant infringement of privacy – not only because one is being "watched" in the (assumed) private sphere of one's home or office, but also because of the increasing pervasiveness of social norms that may arise out of more and more behavioural profiling ("you *should* like this content – 99% of women *do* like these pages!") (43; 20).

4 Challenge 2: The user's ranking depends on the purpose of using that ranking

The problem Consider again the hairdresser-search example, and assume that the user has specified enough details and/or the system knows enough about the (current) location that the top-ranking items are hairdressers in the user's vicinity. This may be (a) good if the user wanted to look up the telephone number or exact address of a hairdresser she already knows, (b) bad if the user wanted to obtain a recommendation for a new hairdresser, or for a type of beauty-care service she wouldn't have thought of otherwise.

This points to two further issues, both showing that the adequacy of a result ranking depends on what one wants to do. First, there are different types of searches, ranging from known-item searches to more open-ended searches. Each of them in turn may be the search for certain *information*, for a *navigational* cue (e.g., searching the homepage of something or someone in order to further explore from there), or for a *transactional* purpose (e.g., searching a site on which to buy something). This classification of search tasks was proposed by Broder (9). Second, each of the tasks again may interact with what type of information one expects: to obtain what one would have thought anyway (= that specific hairdresser in my street), or to be supplied with new and unexpected information. While the latter is a more common expectation in recommender systems, it may also be the expectation in certain usage contexts of a more traditional search engine.

Here, I will address the latter issue as an instance of the purpose of using the ranking, and show that it requires a change in the evaluation criteria. In contrast to this, the search task classification navigational/informational/trans-

actional may be considered to be another context variable, which could be addressed in the ways mentioned in Section 3. The search-task classification will be studied in its role as a mediator variable in Section 6.

The problem is then posed as follows: When one obtains a user ranking U, especially when one gets it from historical data like logfiles, one risks limiting the view to "items the user would have looked at anyway". A system ranking that reflects this U as accurately as possible may be obvious to the user and therefore uninteresting.

Solution approach for evaluation: Beyond accuracy To solve this problem, researchers have proposed to complement the search for and evaluation of accurate system rankings by metrics of desiderata like novelty or serendipity (21).

Novelty means that items that the user is not familiar with should be ranked higher. Evaluation measures can then be adapted in straightforward ways, e.g. by weighting, to reward resulting rankings that have more novel items.

Serendipity means that items that the user might not otherwise have discovered should be ranked higher. Murakami, Mori and Orihara (37) proposed a metric to measure the serendipity of a ranking:

$$unexpectedness = \frac{1}{N}\sum_{i=1}^{N} max(s_i - s_i^{prim}, 0) * isrel(x_i) \tag{8}$$

where s_i^{prim} is the result of a primitive prediction method for the ith item x, s_i is the result of the used prediction method, and N is the number of items ranked. $isrel(x_i) \in \{0, 1\}$ is a function denoting that the item is related to the user's preferences (1) or not (0). To ensure that unexpected items are highly ranked, Murakami et al. proposed another metric *unexpectedness_r*. Let $count(i), (i = 1, \ldots, N)$ denote the number of items suited to the user's preferences lying above the i-th rank in the recommendation list. Then

$$unexpectedness\_r = \frac{1}{N}\sum_{i=1}^{N} max(s_i - s_i^{prim}, 0) * isrel(x_i) * \frac{count(i)}{i}. \tag{9}$$

The result can be calculated overall or, if rankings differ by user, aggregated over users.

Zhang and Hurley (57) addressed the problem of generating recommendation lists that not only contain novel and relevant items, but also exhibit *diversity*. A set is diverse if the inter-item similarities are low. They proposed a method from economics for assessing inequality, based on concentration curves and an associated index, to analyse the bias of recommendation algorithms against the user's novel preferences.

Solution approaches for creating the ranking: use different criteria The novelty of a system-generated ranking can be enhanced by integrating features a into the evaluation of an item x in Equation (1) that are based on metadata like publication date (assuming newer is better), or that take the user's interaction history into account (assuming that a not-yet inspected item is new to the user).

A simple proposal for increasing the serendipity or a ranking is to remove all "obvious" items from S, where obviousness could be overall popularity (e.g. of a movie) in the recommender community, coverage on the Web, etc. Alternatively, an item's score for the current user may be divided by its average score for all users, and the results re-ranked.

As argued above, the importance of different properties of rankings, such as novelty, serendipity or diversity, will differ with the purpose of using that ranking. It may also differ by further user preferences. Therefore, care has to be taken when integrating these modifications into the ranking function.

5 Challenge 3: Querying and ranking is an iterative process

The problem The metrics proposed so far assume a very static process of doing a search or obtaining a recommendation: the user obtains one list, which is more or less good to the extent that it it is accurate, novel, serendipitous, etc. However, this happens only seldom. More usually, some items are inspected by the user, which may change their information state, their preferences, and often also lead them to issue a new query or ask for new recommendations.

Solution approach: Relevance feedback The basic idea of relevance feedback (46) is that the items from a result list that are indeed relevant can serve to refine/modify the query, in the direction of "documents more like this one".

A basic form of this relevance feedback assumes that the query is, like the item (document), textual. Any document $\boldsymbol{x}$ then has a rank $s(\boldsymbol{x}) = rank(\boldsymbol{Q}, \boldsymbol{x})$ with respect to the original query $\boldsymbol{Q}$. Both $\boldsymbol{Q}$ and $\boldsymbol{x}$ are modelled as feature (e.g. word) vectors. After a user has indicated, in step t, some items as relevant (for example by clicking on them or otherwise directing attention to them, see Section 6) and, by exclusion, the others as non-relevant, the updated query is computed as follows:

$$\boldsymbol{Q}_t = \alpha * \boldsymbol{Q} + \beta * \frac{1}{\mid R_t \mid} \sum_{\boldsymbol{x} \in R_t} \boldsymbol{x} + \gamma * \frac{1}{\mid N_t \mid} \sum_{\boldsymbol{x} \in N_t} \boldsymbol{x}, \tag{10}$$

where R_t is the set of relevant items and N_t the set of non-relevant items. $s(\boldsymbol{x})$ is then recomputed by re-ranking, i.e. via $rank(\boldsymbol{Q}_t, \boldsymbol{x})$.

Obviously, there is much room for improvements, for example to avoid solely positive-feedback loops that might compromise the serendipity of the updated result list(s). Many variants of relevance feedback have been proposed, cf. the survey (48).

The repetition of the same or similar queries by different users may be considered another form of iteration. Radlinski and Joachims (44) proposed to learn better rankings from the clickthrough behaviour of previous users with the same information need.

Figure 1: A fictitious scanpath overlaid on a recommender-systems output page (page from `http://movielens.org/html/tour/images/nicepreds.gif`)

6 Challenge 4: The external ranking is only perceived partially

The problem In the previous sections, it has been assumed that the user can see all of S, such that evaluation measures that investigate all of S for suitability are adequate. However, this assumption is generally not warranted.

When a user is confronted with an information presented on a (printed or computer-screen) page, she needs to read it. This requires directing visual attention to the information and – because visual resolution outside the very small foveal area is too low for reading – making an eye movement to fixate on that information. These fixations can be measured by various types of eye-tracking equipment and visualised in overlays on the regarded information. Popular visualisations of single user's eye movements are scanpaths (see Fig. 1), popular visualisations of aggregation of several users' scanpaths are heatmaps (see Fig. 2).

The heatmaps clearly show that – regardless of content – people generally inspect pages in an "F shape". This means that the "organic ranking visibility" is high only for the first few results and then drops off sharply. One example are the percentages of participants that looked at listings in the different ranks reported in Table 1; other studies have produced similar numbers.

In addition, this behavioural tendency of Web users for "information snacking" has increased over the past years, see Fig. 3 based on 1997–2002 data from the metastudy of Lewandowski and Höchstötter (31) and 2005 data from Lorigo

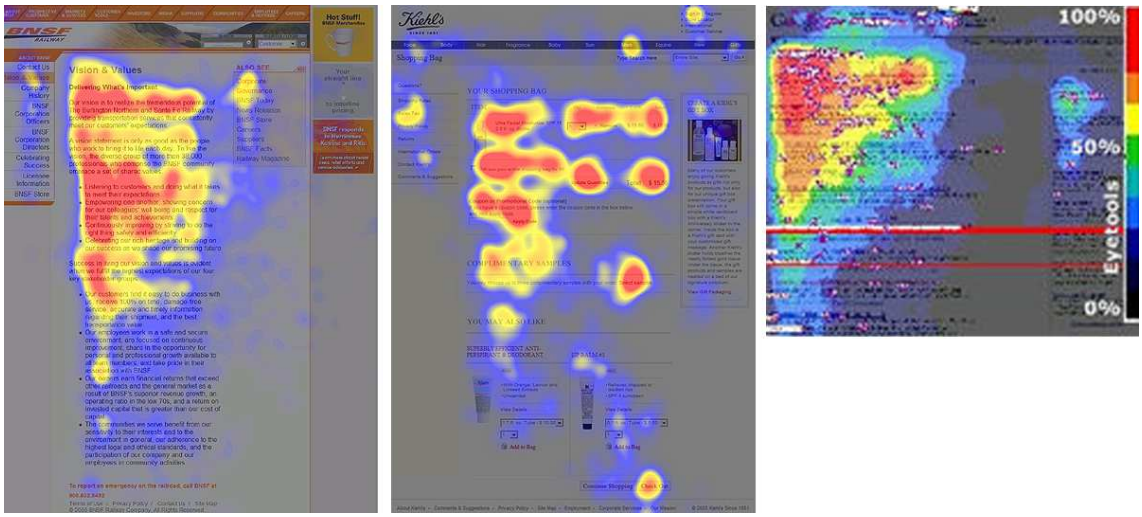

Figure 2: Eye-tracking heatmaps for an "About us" page and a shopping-cart page (38), and a search-engine result page (17). Areas of frequent fixations are shown in red (colour display of this article) / as dark areas bounded by light boundaries (greyscale display).

et al. (33).

While the general "F shape" is an invariant, its details appear to interact with the task. Several studies have investigated the search-task types proposed by Broder (9). They can be illustrated by the task examples used in (33) and (12):

Navigational search tasks

Find the homepage of Michael Jordan, the statistician.
Find the homepage for graduate housing at CMU.

Informational search tasks

Who discovered the first modern antibiotic?
Where is the tallest mountain in NY located?
You are searching for information about loans for a renovation. Select the Web site on which you would like to search information.

Transactional search tasks

You would like to contract a loan for a renovation. Select the Web site on which you would like to contract the loan.

Differences in scanpaths can be visualised by contrasting scanpaths or heatmaps, as in Fig. 4 (which uses height instead of colour), or they can be coded by numerical measures of fixation behaviour, such as

- average numbers of result pages viewed
- average time to complete the task
- average time spent on Web documents / search-engine result pages per question

Rank	Visibility
1	100%
2	100%
3	100%
4	85%
5	60%
6	50%
7	50%
8	30%
9	30%
10	20%

Table 1: Organic ranking visibility (17)

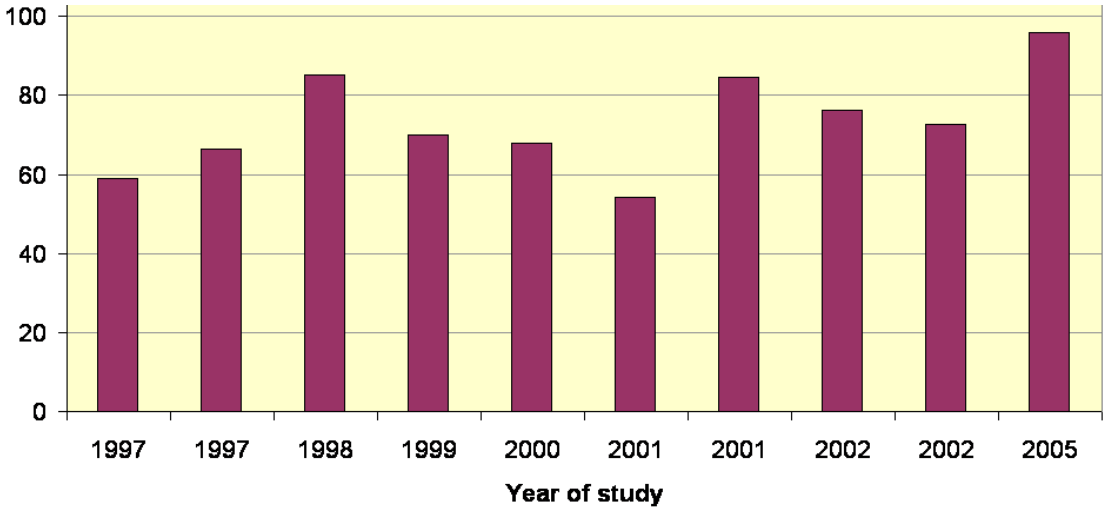

Figure 3: Percentages of users who looked only at the first result screen

- average rank of selected document
- percentage of subjects who clicked on abstract $n, n = 1, ..., 10$
- average number of fixations on a Web document
- average pupil dilation on a Web document over documents selected per question (another indicator of visual attention)
- percentage of repeat viewings of abstracts
- percentage of (strictly) linear scanpaths
- percentage of subjects who fixated on abstract $n, n = 1, ..., 10$

Here, a scanpath is classified as linear or strictly linear as follows (33): Fixations are associated with "regions of interest", numbered by the rank of the fixated item. Thus, for example, the scanpath in Fig. 1 is coded as [2,2,4,1]. A

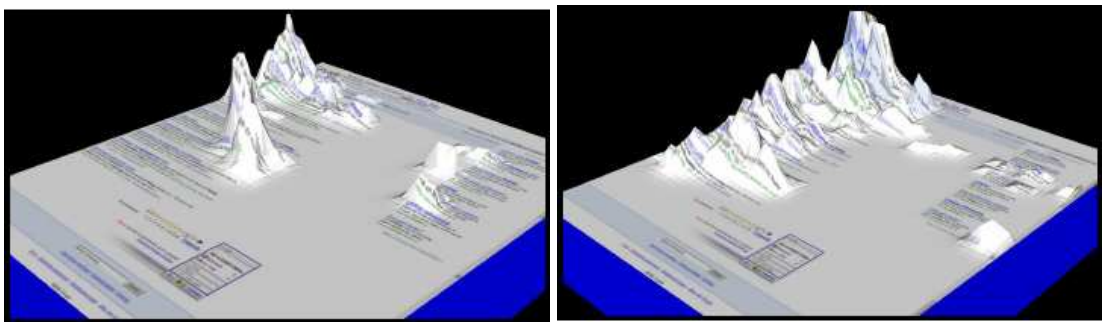

Figure 4: Heatmaps for an informational (left) vs. a transactional (right) search task. Both figures from (12)

linear scanpath is non-decreasing (thus, the whole example is not linear, but its first three fixations are). A strictly linear scanpath is strictly increasing (in the example, only the subpath [2,4] is strictly increasing).

Lorigo et al. (33) compared behaviour on informational and navigational tasks. They found that when participants were given an informational search task, they (a) spent more time to complete the task, (b) spent more time per question on Web documents (but fewer on the search-engine result page), (c) displayed a larger number of fixation and a larger pupil dilation on a Web document relative to the number of documents selected for the question. De Vos & Jansen (2007) compared behaviour on informational and transactional tasks. They found that when participants were given a transactional search task, they (a) looked at a higher number of results and (b) spent slightly more time viewing each result, which led to (c) a longer total time viewing the results. In addition, as Fig. 4 indicates, transactional tasks led to a more systematic exploration of the results page.

Solution approach: search-engine optimisation A pragmatic solution to the problem of limited attention is sought by content providers: they search for the best ways of search-engine optimisation (SEO) to ensure they are displayed at as high a rank as possible. This means either affecting $s()$ (for example, by registering one's site for keywords as in Google Adwords[4]) or changing the attributes a of one's resources to optimise the resulting s value. While most SEO strategies are unreliable and search engines repeatedly emphasise that they strive for "non-manipulable" rankings, SEO has spawned a huge market and has repercussions on, for example, online advertising (56).

Solution approaches: "Top-heavy" evaluation measures Limited attention and the near-exclusive focus on the top-ranked results also affect the perceived quality of the whole ranking. They have led to the development of several evaluation measures that give more weight to matches in the top ranks. A very simple solution is to replace precision by precision@n (see Equation 5), with n small. Mea-

[4] `http://adwords.google.com`

sures that take the rating/ranking into account more fully are the *half-life utility metric HUM* and the *normalised discounted cumulative gain NDCG*.

HUM is defined as

$$HUM = \frac{max(u_i - d, 0)}{2^{(i-1)/(\alpha-1)}} \tag{11}$$

where u_i is the user's rating of item i, d is the default rating, and α is the half-life: the rank of the item on the list such that there is a 50% chance that the user will view that item. *NDCG* at rank position p is defined as

$$NDCG_p = \frac{DCG_p}{IDCG_p}, \text{ where } DCG_p = \sum_{i=1}^{p} \frac{2^{u_i} - 1}{log_2(1+i)} \tag{12}$$

where u_i is the graded relevance of item x_i (i.e. the user's assessment), and $IDCG_p$ is the "ideal" *DCG* at position p, i.e. that obtained when documents of a result list are sorted by *DCG*. In a perfect ranking, *IDCG* and *DCG* are equal.

7 Challenge 5: The platonic S and U don't exist – information systems do!

The problem As the previous challenge has already illustrated, it is not enough to consider abstracted versions of preference lists. Rather, concrete interactions between users and computers have to be taken into account – only together do user(s) and computer(s) form an *information system*. In these interactions, visual attention is not the only factor to be looked at. Rather, as in any software, criteria that enhance (or threaten) usability have to be considered in system design.

Usability is defined as the effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in particular environments (14), which is captured by ISO 9241 (23; 24). *Effectiveness* is the accuracy and completeness with which specified users can achieve specified goals in particular environments. *Efficiency* denotes the resources expended in relation to the accuracy and completeness of goals achieved. *Satisfaction* comprises the comfort and acceptability of the working system. Usability problems remain a key issue on the Web, as shown for example by the repeatedly updated study of "top usability mistakes" by Nielsen (40).

Solution: Usability metrics Important considerations in usability include training and learning how to use a system and reactions to errors. Therefore, a breakdown by "usability objectives" is helpful for operationalising effectiveness, efficiency and satisfaction metrics (14). A comprehensive treatment of usability metrics was given by Tullis and Albert (50).

When tasks are known (which is usually the case in experiments or studies with clear instructions), many basic usability metrics (39) can be easily derived from log data, for example: time needed to complete the task, error rate, number of pages accessed, number of links clicked, number of mouse clicks, distance scrolled, distance cursor moved, or time on each page. Examples of such task analysis are (16; 2).

Additional challenges arise when systems are user-adaptive: They should (a) leave the user in control, (b) make their actions understandable, (c) make their actions predictable, (d) protect privacy, and (e) not limit the users' breadth of experience (25). A current research challenge is the operationalisation of these more involved notions. Standard methods today rely on questionnaires, but these have the drawbacks of reactive methods (the act of questioning may influence the results, only small samples of users can be tested, etc.). Observational measures that can be assessed non-reactively like the ones named above are clearly a desirable addition to these measures.

8 Challenge 6: Preference may depend on framing

The problem The previous sections have assumed that user preferences are well-defined, as long as one knows the full context. However, even when all "objective" features of the items and the usage context can be measured and adequately factored into the system's ranking function, users may still impose different preferences on given sets of items depending on how these are "framed". Framing is the (usually linguistic) embedding of some content into a (often implicit) background structure of meaning.

This could be considered as just another aspect of context, but I treat it as separate because it is often not externally given and pre-existing to the interaction between a user and a system, but rather may be an emergent feature of that very interaction. As such, it becomes an issue for interface design.

Framing and its effect on preferences are best explained with the classical experiment performed by Tversky and Kahnemann (51) as part of the development of their "prospect theory", work that was rewarded with a Nobel Prize in 2002. Tversky and Kahnemann presented the following two decision problems to representative samples of physicians:

Version 1: "Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: If program A is adopted, 200 people will be saved. If program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved. Which of the two programs would you favor?"

Version 2: "Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: If program C is adopted, 400 people will die. If program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die. Which of the two programs would you favor?"

While the described option pairs A–C and B–D are equal from a probabilistic standpoint, they are framed positively (in terms of living and saving people, or

gains) in version 1, and negatively (in terms of dying and losses) in version 2. In the positive frame, over 70% of respondents chose the safe option A, the rest chose the risky option B. In the negative frame, over 70% of respondents chose the risky option D, only just over 20% chose the safe option C.

In (4), we observed similar phenomena in a Web recommendation context. In an experimental online store, an anthropomorphic agent posed 56 questions in a sales dialogue, ostensibly to produce a recommendation ranking. In normal sales dialogues, customers lose patience far before 10 questions have been asked, and they are very critical of "inappropriate" questions. In the study, 35-50% of the questions were non-legitimate and/or irrelevant, which had been established in an independent prior empirical study. Also, most participants described themselves as privacy-conscious. Yet, in spite of the large number of questions and the inappropriateness of most of them, 54% of participants answered at least 98% of the questions, although they had previously agreed to the sale and further usage of their data. They reported a large satisfaction with the process and did not report any privacy concerns arising from the interaction.

Upon closer inspection of these surprising results and the underlying materials, we found that in the pre-test of the questions, these had been framed in terms of a *loss* of privacy, while inside the store, the *gains* in personalisation to be had from answering the questions were emphasised. We concluded that, together with strong tendencies to interpret the conversation with the agent as requiring cooperative behaviour, these framing effects made irrelevant questions seem relevant and non-legitimate questions seem legitimate.

A similar interpretation can be made of the findings of Kobsa and Teltzrow (27). They observed that online shoppers felt well-supported in their book choices by a list of recommendations that was in fact constant, and that they were willing to disclose personal data in return for this seeming personalisation. This suggests that framing a ranked recommendation list as arising from data mining creates the impression of personalisation.

Solution approaches? Framing is a well-known problem in the behavioural sciences, and it has been under much scrutiny in media analyses, see for example (49). Recent work calls for raising more awareness of the dangers of framing in the general public (29). However, to the best of our knowledge, such efforts are only beginning to emerge with respect to the Web.

With regard to document search, framing effects pose a specific challenge: If ranking is based only on the "factual content" of documents, it will miss out on an important source of human preferences of documents over others – recall that in the example, the "factual content" of options A and C was identical, as was that of B and D, but A was preferred over C and D over B. If, on the other hand, document-processing techniques such as natural-language understanding were able to extract framing, how should the system deal with it when producing the ranking? Should it try to comply with the user's favourite framing, or framing-induced preferences? Or should it try to counteract them, for example by focussing on presenting a diversity of framings? We believe that successful solution approaches should, first and foremost, help people become more aware of the presence of framing and its influences.

9 Challenge 7: System-use dynamics and attitudes

The problem Information retrieval and Web search have become such an integral part of everyday activities that there is a strong tendency to regard search-engine results as "the truth". This is particularly true of Google, which continues to claim major market shares in most countries. It is therefore likely that Web users will trust these programs and assume that their ranking is a "true" ordering of relevance.

To investigate this possibility, Pan et al. (42) investigated eye movements and other behavioural and self-report measures of users who were exposed to one of three experimental conditions: Upon entering a search query, they received either the ranking returned by Google, the ranking returned by Google with the first and second results exchanged, or the reversed Google ranking. In all three conditions, the interface was the familiar Google result list layout. In all three conditions, users displayed the typical behaviour of viewing mainly the first two or three results (see Section 6), with the same number of fixations on the first two results in all three conditions.

Their clicking behaviour showed an even more strongly pronounced preference for the first and second results in the first two conditions. In the "reversed" condition, the first and second results were also clicked on more often than the remaining ones, although the drop-off was not as pronounced. Also, participants scrutinised the whole result page for longer. This may indicate that they did notice that something was unusual, or even wrong, about this seemingly innocuous ranking; however, when questioned later about their search experience, users attributed the problems to themselves, stating that they "did not have much luck with several of the questions" or that they "could not think of the right search terms".

These findings may be interpreted as another instance of framing: presenting a ranking as generated by an otherwise trusted search engine incites people to believe it is a good ranking. This may be considered as originating from trust in the "brands" that the big search engines have built (26). In contrast to the results mentioned in Section 8, however, this framing is generated by the dynamics of using Web technology, and this framing changes the basic evaluation framework used throughout the previous sections: S appears to co-determine U, such that a comparison between the two rankings is not meaningful any more as a measure of S's quality in approaching a user ranking, but instead as a measure of S's quality in influencing it.

Solution approaches? This problem is an instance of the more general observation that measuring something often influences it.[5] It calls for a more careful and wider view of evaluation as such: from a "natural-science" style of evaluation in which researchers regard themselves and their procedures as objective observers outside the process, to a more "social-science" style of evaluation in which researchers regard themselves and their procedures as participants of the

[5] Butler (10) provided a demonstration of this phenomenon in the context of ranking universities: As soon as Australian universities started being evaluated according to the number of their publications, these numbers rose steeply.

process. While this problem is well-known in general evaluation theory and has been addressed by various methods, it has, to the best of our knowledge, not been attended to in evaluation frameworks for Web search and recommendation.

10 Challenge 8: People have mistaken beliefs about ranking functions

The problem Things can be ranked by many criteria, but – even if one knows all the ingredients of $s()$ and its underlying functions – what does this ranking really mean? Often, an overall view of "good" / "better" is taken, and the details of context (good/better for what?) are forgotten. To the extent that the details of the $s()$ function are known and discussed, some attributes $a()$ or valuations of them $v()$ are generally believed to have some properties, and these properties are believed to contribute to the overall "goodness". However, these beliefs may be wrong and due to an overpopularization of rankings and an associated use of them for purposes they are not adequate for.

On the Web, all the details of attributes and valuations are subject to continuing efforts by Web-content providers at search-engine optimisation, and continuing efforts by search-engine providers at making their criteria and algorithms ever more "non-manipulable" indicators of quality. To the extent that search engines and search-engine optimisation are commercial, the details of criteria and ranking functions remain largely proprietary. In other application areas of ranking, criteria are – often – better known publicly. A case in point are bibliometric rankings, which are increasingly being used to rank scientists and make decisions like appointments based on these rankings.

It is likely that bibliometric rankings are subject to many of the same challenges as search-engine rankings – the dependency of "true" quality/adequacy rankings on context and purpose, the iterativeness and system dynamics of rankings, and the implications of human perceptual and cognitive biases and of human-computer interaction. In addition, the long history of bibliometrics comprises a rich tradition of inspecting the measures used with respect to further "ground truths". In many instances, this research has shown that popular beliefs about effects are not true or that even the reverse holds. For example, van den Besselaar and Leydesdorff (52) showed that scientists who received most funding had experienced the most rejections and worst performance in the past and vice versa – contrary to public belief which would expect the opposite. It is also often found that outside the top- and bottom-ranked authors, rankings do not correlate much with other indicators of quality and success, see for example (52; 18).

Glänzel (19) took a more general look at this problem and identified "Seven Myths in Bibliometrics". He described, as the first of three classes of myths, "myths that reflect dreams and visions, and are used as excuse for unsatisfactory results, or serve as recipe for hoped-for-success", such as "collaboration is always a guarantee for success" or "citations are measures of scientific quality". Another class are "myths that are fostered by mistrust", such as "self-citations are very harmful and must be removed from the statistics" and "reviews are inflating impact". The third class are "myths that have their roots in uninformed use of

data, in misunderstandings or ignorance" such as the belief that "averages must not be used in bibliometrics". Drawing on a wide range of data analyses and theoretical arguments, he showed that these beliefs are, as general statements, false – thus, myths.

Solution approaches? These misunderstandings call, above all else, for more education of decision makers and the general public about what ranking and ranking criteria mean, how they work, and what their limitations are. Scientist are called upon in their role as experts, in order to avoid acting as (even if unwilling) "catalyst[s] in the process of fostering, disseminating and extending these myths" (19). Decision makers who determine rankings should become sensitive to having to use multiple criteria and maximal transparency; they may even profit from the recommendation made by Pan et al. (42) for the context of search engines: "... Users might benefit from having more information regarding the mechanisms by which search engines ... [rank] [and from search-engine design showing search-engine limitations]. ... [A] certain degree of randomness in the ranking of returned results ... leads to improved search."

11 Conclusions and outlook

In this paper, I have argued that to evaluate the quality of a ranking such as done in search engines or recommender systems, one must answer the question whether the ranking is the "right ranking" for the given person, in the given circumstances. I have proposed a very simple general conceptualisation of the ranking-evaluation task: the comparison between the ranking generated by a computational system such as a search engine, and the "true ranking inside the user's head". The article then proceeded to describe eight challenges to this simple model. By investigating the *problems* posed by the realities of human cognition and behaviour in human-machine interaction, and by describing popular *solution approaches* to these problems, a number of answers to the question have been obtained – or methods for obtaining such answers.

For the later challenges in particular, there are currently hardly any solutions. The reason may be that they require a highly interdisciplinary approach and, in some cases, radical re-thinkings of how the mathematics of ranking can or should be used. Nonetheless, I hope that by giving this exemplary overview of a wide range of research from diverse areas, and by pointing to places where straightforward solutions are bound to fail, to have provided some inspiration for addressing some of the problems in novel ways.

This paper has – by design – been exemplary, both in terms of the literature cited and in terms of the problems raised. In future work, we will address further relevant issues for ranking and its evaluation. One research direction will be to enrich the evaluation-centric framework used here, which treats the derivation of system rankings mainly as a black box, by a more detailed investigation of the document-query matching process (7). In addition, we intend to further investigate the proposals made in (5) to address the problem that user preferences are probably random variables, that their mental rankings may only be partial or-

ders, and that the nature of preferences as well as how recommendations should be made may change substantially when moving from the idealised individual user to users embedded in their real-life social groups.

References

[1] A. Altman and M. Tennenholtz. Axiomatic foundations for ranking systems. *Journal of Artificial Intelligence Research*, 31:473–495, 2008. `http://www.jair.org/media/2306/live-2306-3748-jair.pdf`, retrieved 2009-06-15.

[2] Richard Atterer and Albrecht Schmidt. Tracking the interaction of users with ajax applications for usability testing. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1347–1350, New York, NY, USA, 2007. ACM.

[3] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[4] B. Berendt, O. Günther, and S. Spiekermann. Privacy in e-commerce: Stated preferences vs. actual behavior. *Communications of the ACM*, 48(4):101–106, 2005.

[5] Bettina Berendt and Veit Köppen. Improving information ranking by respecting the multidimensionality and uncertainty of user preferences. In Giuliano Armano, Marco de Gemmis, Giovanni Semeraro, and Eloisa Vargiu, editors, *Intelligent Information Access*, Studies in Computational Intelligence, Berlin etc., 2009. Springer.

[6] Bettina Berendt and Anett Kralisch. A user-centric approach to identifying best deployment strategies for language tools: The impact of content and access language on web user behaviour and attitudes. *Journal of Information Retrieval*, 12(3):380–399, 2009.

[7] David Bodoff and Stephen E. Robertson. A new unified probabilistic model. *Journal of the American Society for Information Science and Technology*, 55(6):471–487, 2004.

[8] Pia Borlund and Ian Ruthven. Introduction to the special issue on evaluating interactive information retrieval systems. *Information Processing & Management*, 44(1):1–3, 2008.

[9] Andrei Z. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[10] L. Butler. A list of published papers is no measure of value. *Nature*, 419:877, 2002.

[11] S. Chakrabarti. *Mining the Web*. Morgan Kaufmann, San Francisco, CA, 2003.

[12] de Vos & Jansen. Visual attention to online search engine results, 2007. `http://www.checkit.nl/pdf/eyetracking_research.pdf`, retrieved 2009-03-03.

[13] Anind K. Dey, Gerd Kortuem, David R. Morse, and Albrecht Schmidt. Editoral: Situated interaction and context-aware computing. *Personal and Ubiquitous Computing*, 5(1):1–3, 2001.

[14] Alan Dix, Janet Finlay, Gregory Abowd, and Russell Beale. *Human Computer Interaction*. Prentice Hall Europe, 1998.

[15] Leo Egghe. New relations between similarity measures for vectors based on vector norms. *Journal of the American Society for Information Science and Technology*, 60(2):232–239, 2009.

[16] Michael Etgen and Judy Cantor. What does getting WET (web event-logging tool) mean for web usability? In *Fifth Human Factors and the Web Conference*, 1999. http://zing.ncsl.nist.gov/hfweb/proceedings/etgen-cantor/index.html, retrieved 2009-06-15.

[17] Eyetools. Eyetools research and reports: Eyetools, enquiro, and did-it uncover search's golden triangle, 2008. http://www.eyetools.com/inpage/research_google_eyetracking_heatmap.htm, retrieved 2009-06-15.

[18] Wolfgang Glänzel. On the h-index – a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67(2):315–321, 2006.

[19] Wolfgang Glänzel. Seven myths in bibliometrics. about facts and fiction in quantitative science studies. *Collnet Journal of Scientometrics and Information Management*, 2(1):9–17, 2008. Conference version at http://www.collnet.de/Berlin-2008/GlanzelWIS2008smb.pdf, retrieved 2009-06-15.

[20] S.F. Gürses, B. Berendt, and Th. Santen. Multilateral security requirements analysis for preserving privacy in ubiquitous environments. In *Proceedings of the Workshop on Ubiquitous Knowledge Discovery for Users at ECML/PKDD 2006*, pages 51–64, Berlin, September 2006. http://vasarely.wiwi.hu-berlin.de/UKDU06/Proceedings/UKDU06-proceedings.pdf.

[21] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[22] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. Demographic prediction based on user's browsing behavior. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 151–160. ACM, 2007.

[23] International Organization for Standardization. ISO 9241-400:2007. ergonomics of human–system interaction – part 400: Principles and requirements for physical input devices, 2007. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38896, retrieved 2009-06-15.

[24] International Organization for Standardization. ISO 9241-151:2008. ergonomics of human-system interaction – part 151: Guidance on world wide web user interfaces, 2008. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=37031, retrieved 2009-06-15.

[25] Anthony Jameson. Adaptive interfaces and agents. In Julie A. Jacko and Andrew Sears, editors, *Human-Computer Interaction Handbook*, pages 305–330. Erlbaum, Mahwah, NJ, 2003. http://dfki.de/~jameson/abs/Jameson03Handbook.html, retrieved 2009-06-15.

[26] Bernard J. Jansen, Mimi Zhang, and Ying Zhang. Brand awareness and the evaluation of search results. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1139–1140, New York, NY, USA, 2007. ACM.

[27] A. Kobsa and M. Teltzrow. Impacts of contextualized communication of privacy practices and personalization benefits on purchase behavior and perceived quality of recommendation. In *Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research (IUI 2005)*, pages 48–53, San Diego, CA, 2005.

[28] Anett Kralisch and Bettina Berendt. Language-sensitive search behaviour and the role of domain knowledge. *The New Review of Hypermedia and Multimedia*, 11(2):221–246, 2005.

[29] George Lakoff. *Don't think of an elephant! Know your values and frame the debate*. Scribe Publications, Carlton North, Vic., Australia, 2005.

[30] Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[31] Dirk Lewandowski and Nadine Höchstötter. Web searching: A quality measurement perspective. In Amanda Spink and Michael Zimmer, editors, *Web Searching: Interdisciplinary Perspectives*. Springer, Dordrecht, The Netherlands, 2007.

[32] Hugo Liu and Rada Mihalcea. Of men, women, and computers: Data-driven gender modeling for improved user interfaces. In *Proceedings of the International Conference on Weblogs Social Media (ICWSM)*, pages 121–128, 2007.

[33] Lori Lorigo, Bing Pan, Helene Hembrooke, Thorsten Joachims, Laura A. Granka, and Geri Gay. The influence of task and gender on search and evaluation behavior using google. *Inf. Process. Manage.*, 42(4):1123–1131, 2006.

[34] Thierry Marchant. An axiomatic characterization of the ranking based on the *h*-index and some other bibliometric rankings of authors. *Scientometrics 80(2), 325342.*, 2009. http://www.springerlink.com/content/e71h95u774701j1k, retrieved 2009-06-15.

[35] MaxMind, Inc. GeoIP city accuracy for selected countries, 2008. http://www.maxmind.com/app/city_accuracy, retrieved 2009-06-15.

[36] Mozilla Labs. Introducing geode, 2008. http://labs.mozilla.com/2008/10/introducing-geode/, retrieved 2009-06-15.

[37] Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. Metrics for evaluating the serendipity of recommendation lists. In *New Frontiers in Artificial*

Intelligence, JSAI 2007 Conference and Workshops, Miyazaki, Japan, June 18-22, 2007, Revised Selected Papers, volume 4914 of *Lecture Notes in Computer Science*, pages 40–46. Springer, 2008.

[38] J. Nielsen. Eyetracking research, 2006. `http://www.useit.com/eyetracking`, retrieved 2009-06-15.

[39] Jakob Nielsen. Usability metrics: Tracking interface improvements. *IEEE Software*, 13(6):12–13, 1996.

[40] Jakob Nielsen. Top ten mistakes in web design, 2007. `http://www.useit.com/alertbox/9605.html`, retrieved 2009-06-15.

[41] Jakob Nielsen and Kara Pernice. *Eyetracking Web Usability (Voices That Matter)*. Addison-Wesley Longman, Amsterdam, 2009.

[42] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. In google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.

[43] D.J. Phillips. Privacy policy and PETs: The influence of policy regimes on the development and social implications of privacy enhancing technologies. *New Media Society*, 6(6):691–706, 2004.

[44] Filip Radlinski and Thorsten Joachims. Query chains: learning to rank from implicit feedback. In Robert Grossman, Roberto J. Bayardo, and Kristin P. Bennett, editors, *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 239–248. ACM, 2005.

[45] Stephen Robertson and Hugo Zaragoza. On rank-based effectiveness measures and optimization. *Inf. Retr.*, 10(3):321–339, 2007.

[46] J. J. Jr. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Inc., 1971.

[47] Ronald Rousseau. Woeginger's axiomatisation of the h-index and its relation to the g-index, the $h^{(2)}$-index and the R^2-index. *Journal of Informetrics*, 2(4):335–340, 2008.

[48] Ian Ruthven and Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2):95–145, 2003.

[49] H.A. Semetko and P.M. Valkenburg. Framing European politics: a content analysis of press and television news. *Journal of Communication*, 50(2):93–109, 2000.

[50] Thomas Tullis and William Albert. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann, 2008.

[51] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211:453–458, 1981.

[52] Peter van den Besselaar and Loet Leydesdorff. Past performance as predictor of successful grant applications: A case study. SciSA Report 0706, Rathenau Institute, The Hague, December 2007. http://home.medewerker.uva.nl/p.a.a.vandenbesselaar/bestanden/2007%20magw.pdf, retrieved 2009-06-15.

[53] Alex Varshavsky, Eyal de Lara, Jeffrey Hightower, Anthony LaMarca, and Veljo Otsason. GSM indoor localization. *Pervasive and Mobile Computing*, 3(6):698–720, 2007.

[54] Liwen Vaughan. New measurements for search engine evaluation proposed and tested. *Information Processing & Management*, 40(4):677–691, 2004.

[55] William Webber, Alistair Moffat, Justin Zobel, and Tetsuya Sakai. Precision-at-ten considered redundant. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *SIGIR*, pages 695–696. ACM, 2008.

[56] Bo Xing and Zhangxi Lin. The impact of search engine optimization on online advertising market. In *ICEC '06: Proceedings of the 8th international conference on Electronic commerce*, pages 519–529, New York, NY, USA, 2006. ACM.

[57] Mi Zhang and Neil Hurley. Statistical modeling of diversity in top-n recommender systems. In *Proceedings of the ACM Web Intelligence Conference WI09, pp. 490497. IEEE Computer Society.*, 2009.