Relevance model

Table of Contents

List of figures

List of tables

1 Introduction

The goal of the relevance model is to enhance the ranking performance of the baseline retrieval system by integrating information from multiple sources. The problem to combine information from multiple sources to get an overall ranking of items occurs in several fields, like multi criteria decision making (MCDM), data fusion, meta search, etc.

To get the overall score for an item x, which is represented by the criteria values v_i a common approach is to use the weighted sum or the weighted average of the individual criteria values,

$$w(x) = \sum_i \alpha_i v_i,$$

where the α_i are the weights. In general, the weights must be seen as trade-offs, i.e. a decrease in criterion i can be compensated by an increase of α_j/α_i in criterion j. If the criteria are expressed in equivalent units, the weights might be interpreted as relative importance of the criteria.

The use of the weighted sum is based on the assumption of preference independence. Consider we have two items and the first item is ranked higher than the second. Then, varying the value of a criterion in both items in the same way does not change the relative ranking of the two items (Bouyssou, 2000).

Another important assumption of the weighted sum is that the value scales are linear. I.e. equal differences in the value scales have the same effect independent of the position on the value scale. For example, the difference in the overall score is the same for two documents with 0 and 50 downloads as is the difference for two documents with 500 and 550 downloads.

According to Bouyssou (2000), there are two main approaches (in decision theory) to account for the difficulties in the weighted sum approach:

1. Trying to transform the criteria values so they better fulfill the assumptions of the weighted sum. Especially, trying to linearize the criteria scales, so the effect of differences in the linearized scales is independent of the position on the scale.
2. Instead of comparing the criteria of one alternative and then derive a score for the alternative, we could compare the alternatives pairwise (with respect to all criteria) and then derive a global preference relation from those pairwise preferences.

We will discuss the first approach in the section on the "Scale construction step". The second approach leads to the outranking approach, which has been used in Farah & Vanderpooten (2006) to build a multi criteria information retrieval system inspired by decision theory. The outranking approach is based on the notion of pairwise preference of the alternatives, which directly allows incorporating knowledge about the imprecision in the data as well as indifference statements about the criteria values. This allows a more direct way to model how the criteria interact and therefore, how and when criteria can compensate each other. Further, the pairwise preferences of alternatives build on the pairwise comparison of the criteria values. Hence only entities of the same kind are compared and therefore the different criteria do not have to be expressed on a common scale.

In the first approach, the problem to combine information from multiple sources can be decomposed into two steps: the aggregation step and the scale construction step. One additional step is needed to take care of criteria with heterogeneous coverage.

1.1 Aggregation step

In the aggregation step we need to aggregate two groups of criteria: query dependent and query independent criteria. As mentioned in the introduction, a common approach to aggregate the scores of multiple criteria is to use the weighted average operator (weighted sum). When using the weighted sum both criteria groups must be comparable and therefore be expressed on the same scale.

Further, the weighted sum implements a compensatory logic and is based on the assumption that the criteria are independent. Both properties have been criticized for not being realistic(e.g. da Costa Pereira, Dragoni, & Pasi, 2012; Eickhoff, de Vries, & Collins-Thompson, 2013; Moulahi, Tamine, & Yahia, 2014; Silva et al., 2009). For example, if we assume that topical relevance is a necessary condition for a document being considered relevant (see e.g. Wang & Soergel, 1998), it follows that a lack of topical relevance should not be compensated by a surplus of other criteria.

Da Costa Pereira et al. (2012) proposed an interesting alternative aggregation operator inspired by the ordered weighted average (OWA) operator (Yager, 1988). They propose a prioritized ordered weighted average operator, which is defined by

$$\mathrm{F_s(x)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \lambda_\mathrm{i} \cdot \mathrm{C_i}$$

with $\lambda_\mathrm{i} = \lambda_{\mathrm{i-1}} \cdot \mathrm{C_{i-1}}$ and $\lambda_1 = 1$. $\mathrm{C_i}$ represents the degree of satisfaction of criterion i (therefore $\mathrm{C_i} \in [0,1]$) and the $\mathrm{C_i}$ are ordered by a given priority order of the criteria $C_1 \succ C_1 \succ \cdots \succ C_n$, where $\mathrm{C_1}$ is the most important criterion. The prioritized OWA operator models dependencies between the criteria since the weights depend on the degree of satisfaction of the more prioritized criteria. Hence, the degree of compensation is reduced for lower prioritized criteria depending on the scores of higher prioritized criteria. Note that for the prioritized OWA, one does not have to find a weighting scheme as for the weighted sum.

Further, Silva & Moura (2009) used genetic programming to find optimal aggregation functions composed from a simple set of arithmetic operations. They found that modeling the dependence of textual evidence and PageRank was crucial for navigational queries in a web page retrieval setting.

In the first version of the relevance model we will use the following aggregation function

$$\mathrm{RSV(x)} = \mathrm{score_q} \cdot (1 + \alpha_\mathrm{qi} \cdot \sum \alpha_\mathrm{i} \cdot \mathrm{v_i}),$$

where $\mathrm{score_q}$ is the Lucene/Solr score for the query q, α_qi gives the relative importance of the query dependent criteria over the query independent criteria, and the α_i and $\mathrm{v_i}$ are the weights and scores for the query independent criteria. This aggregation functions avoids the need to normalize the query dependent scores and further models a dependence between query dependent and query independent criteria.

For the query independent criteria we have the standard weighted sum. Therefore, a lack in one criterion can be totally compensated by other criteria. The parameter α_{qi} enables us to tune the relative importance of both criteria groups, although it could of course be incorporated in the α_i.

Note that this aggregation function is actually equivalent to the prioritized OWA operator F_s,if we set $\alpha_{qi} = 1$ and if we define a meta-criterion $C_2 = \sum \alpha_i \cdot v_i$, i.e. all query independent criteria have the same priority. Now let $C_1 = \text{score}_q$ and $\lambda_1 = 1$, then we get $\lambda_2 = \lambda_1 \cdot C_1 = \text{score}_q$ and

$$\text{RSV}(x) = \text{score}_q \cdot (1 + C_2) = C_1 + \lambda_2 \cdot C_2.$$

Note that we have neglected to normalize the Lucene/Solr scores to the unit interval, since a multiplicative factor has no effect on the overall ranking.

1.2 Scale construction step

Scale construction is a difficult and more or less subjective process. In information retrieval (IR) mainly two approaches have been followed in prior works. First, the use of heuristics, and second, learning of utility functions from training data with relevance judgments. In MCDM the most common approach is to directly elicit the utility functions from the decision maker. The most prominent case for which transformed scales have been developed is for the term frequency (tf) in tf-idf based IR systems, since it was observed that raw term frequency overrates higher term frequencies with respect to probability of relevance.

1.2.1 Heuristics

Common tf normalization schemes use heuristics like e.g. $\log(\text{tf})$ (Craswell & Robertson, 2005; Robertson & Walker, 1994; Zhao & Hu, 2014) or $\text{tf}^{1/2}$ (e.g. used in Lucene/Solr) to reduce the effect of higher tf values. Richardson, Prakash, & Brill (2006) used the logarithm of the number of page views, which follows a power law distribution, "to reduce the dynamic range". The BM25 term frequency normalization

$$\text{tfn} = \frac{\text{tf}}{\text{tf} + k}$$

has been derived as an approximation to the 2-Poisson model (Robertson & Walker, 1994). Craswell & Robertson (2005) used the generalized form

$$w(x) = \frac{x^\alpha}{x^\alpha + k^\alpha}$$

as a heuristic for the utility of PageRank scores. This function is a sigmoid of $\log(x)$. The sigmoid has also been proposed by Nottelmann & Fuhr (2003) as an approximation to the probability of relevance for the retrieval status values of an IR system. The intuition thereby is that in the ideal case all r relevant documents are ranked at the first r positions, followed by the non-relevant documents. Then the probability of relevance is given by a step function. The sigmoid is a smoothed version of this step function for the non-ideal case. In both studies the parameters for the heuristics has been learned from training data with relevance judgments.

1.2.2 Utility functions in multi-criteria decision making

As mentioned, in MCDM scales are in general constructed by eliciting preferences from the decision maker. In the Analytic Hierarchy Process (AHP) (Saaty, 1977), scales are constructed from pairwise preference statements given as relative importance estimates between two alternatives with respect to one criterion on a five or nine point scale. In the additive multi-attribute value model the method of *standard sequences* could be used to elicit a value function based on indifference statements (Bouyssou, 2000). This method is based on the independence assumption for the criteria. Suppose we have two alternatives that only differ in one criterion. We could then ask the decision maker for which value of another criterion he gets indifferent between the two alternatives. This process can be iterated to construct a value function over the whole range of values for the other criterion. For details we refer to Bouyssou (2000).

These methods might also be applicable to the problem to find utility functions for the criteria in a relevance model. Though, it is stated that no single ranking can perform optimally in different user situations (e.g. da Costa Pereira et al., 2012). Therefore, one might expect a large variation in elicited utility functions. Though it is not as clear if the variances in the relevance model for different user situations are due to differences in the aggregation (e.g. due to differences in the importance of the criteria) or due to differences in the utility functions. Hence, if we assume that the differences are due to the aggregation then utility functions elicited from a small group of users could be sufficiently stable. Under these circumstances elicited utility functions could be expected to result in better performance than generic heuristics.

1.2.3 Characteristic Scores and Scales

Glanzel & Schubert (1988) proposed the method Characteristic Scores and Scales (CSS) to find a characteristic partition for highly left skewed distributions. They analyzed citation count distributions and used the method to find classes of papers which they interpreted as "poorly cited", "fairly cited", "remarkably cited", or "outstandingly cited".

The classes are constructed in the following way: The first class boundary is set to the mean of the distribution, $\beta_1 = \mu$. Then the distribution is truncated at the first boundary and the second boundary is found at the mean of the truncated distribution, $\beta_2 = \mathrm{mean}(\{x_i | x_i \geq \beta_1\})$. Finally, the k-th class boundary is given by

$$\beta_k = \mathrm{mean}(\{x_i | x_i \geq \beta_{k-1}\}).$$

The iteration should be stopped at some reasonable k_{max},, e.g. if the proportion of elements in class k drops below some threshold. We then set $\beta_{k_{max}} = \sup(x_i)$. Though, the determination of the threshold and therefore the number of classes k_{max} is, to some degree, dependent on (the shape of) the distribution.

The proportions of the classes have been found to be remarkably stable, e.g. for the citation count distributions of different disciplines (Glänzel, 2007). Further, it has been shown that the class boundaries are related to the h-index of the distribution via

$$\beta_k = \left(\frac{A}{h^\alpha}\right)^k$$

for Lotkaian distributions, where A is the total number of citations and α is the parameter of the Lotkaian distribution (Egghe, 2010). Mitzenmacher (2004) noted that such an exponential partitioning has actually been used to derive the Pareto distribution. Though, for empirical distributions this relation will of course only hold approximately at most.

Now suppose we have two samples which are highly left skewed. By applying the CSS method on both samples we get an alignment of the two sample scales. Since we expect, that the characteristic classes contain similar elements, we can state that the class boundaries $\beta_k^{(1)}$ and $\beta_k^{(2)}$ can be seen as equivalent values on the original scales, and the sample scales are therefore aligned according to the CSS class boundaries. There are three distinct class boundaries - the lowest boundary $\beta_0 = 0$, the first boundary $\beta_1 = \mu$, and the largest boundary $\beta_{k_{max}} = \sup(x_i)$. So, the minimal element (zero), the mean and the maximal element of the two samples get aligned. This demonstrates that in general, this alignment is non-linear, which would not be possible for simple linear normalization schemes.

Though, to make use of the CSS method for the relevance model, we must map the CSS classes to scores that we can use in the aggregation. If no prior knowledge is available, we could simply map the classes to equidistant points on a score scale, i.e.

$$\text{score}(x_i) = \frac{k}{k_{max}} \qquad \text{for } x_i \in [\beta_k, \beta_{k+1}).$$

Though, we will linearly interpolate the scores between the class boundaries, to get a continuous transformation. Note that for a Lotka distribution, we simply get an approximation to the logarithm of the original values, $\text{score}_{\text{Lotka}}(x_i) \approx \log(x_i)$.

The mapping of the CSS classes plays a similar role as the utility functions. The mapping can be tuned to better capture the "real" utility of the users. The determination of a customized mapping might be facilitated compared to elicit a utility function due to the lower number of classes. Further, the mapping can be more intuitively based on domain knowledge compared to general tuning parameters as for example, for the sigmoid.

1.3 Heterogeneous coverage of criteria

Most of the criteria are only defined on subsets of the documents, e.g. only on specific document types, or they are only available for specific collections. If we have a set of criteria that are disjoint but their union covers (almost) the total set of documents, we can simply combine these criteria to a meta-criterion.

If a (meta-) criterion does not cover the total set of documents, the uncovered documents would be implicitly demoted (if missing values would be implicitly replaced by zero). Missing values are a common problem in statistics and a standard approach is to replace missing values with the mean of the distribution. Though, for some criteria we cannot differentiate between zero and missing values

(e.g. for citation counts from CitEc[1]). We then prefer to replace missing values with zeros, since otherwise documents with values below average would be demoted only for having a value.

If two criteria, which measure the same entities, but are given on different but partly overlapping subsets (e.g. number of downloads from LogEc[2] and eTracker), replacing the missing values in the complement of the union by the mean would introduce unnecessary imprecision. In this case we prefer to merge both criteria into one criterion by using the original values in the complement and the maximum of the values in the union. We note that we use the maximum value (e.g. instead of the mean) for practical reasons.

2 Relevance measures

In the section "Relevance model" we will give an overview of all the identified relevance factors and how these will be integrated as relevance measures into the relevance model. In this section we will discuss some selected relevance measures in more detail and exemplify the application of the CSS methods to the corresponding data.

2.1 Usage data

We have collected usage data from EconBiz and LogEc. The EconBiz data is collected via the eTracker Event API via a client-side JavaScript function, while the LogEc data contain the aggregated usage data of multiple systems that provide access to the RePEc collection extracted from the corresponding server logs.

As shown in Figure 1, the usage data roughly follow a power law distribution. Though, the EconBiz data exhibits a stronger deviation from the power law for smaller number of downloads. The EconBiz sample differs from the LogEc sample in two distinct ways: First, the EconBiz collection covers many more distinct documents and therefore the downloads can be distributed over many more documents, and second, the EconBiz sample is significantly smaller than the LogEc sample. Therefore the probability to observe a higher number of downloads for a specific document is much more reduced compared to the LogEc sample.

[1] http://citec.repec.org/
[2] http://logec.repec.org/

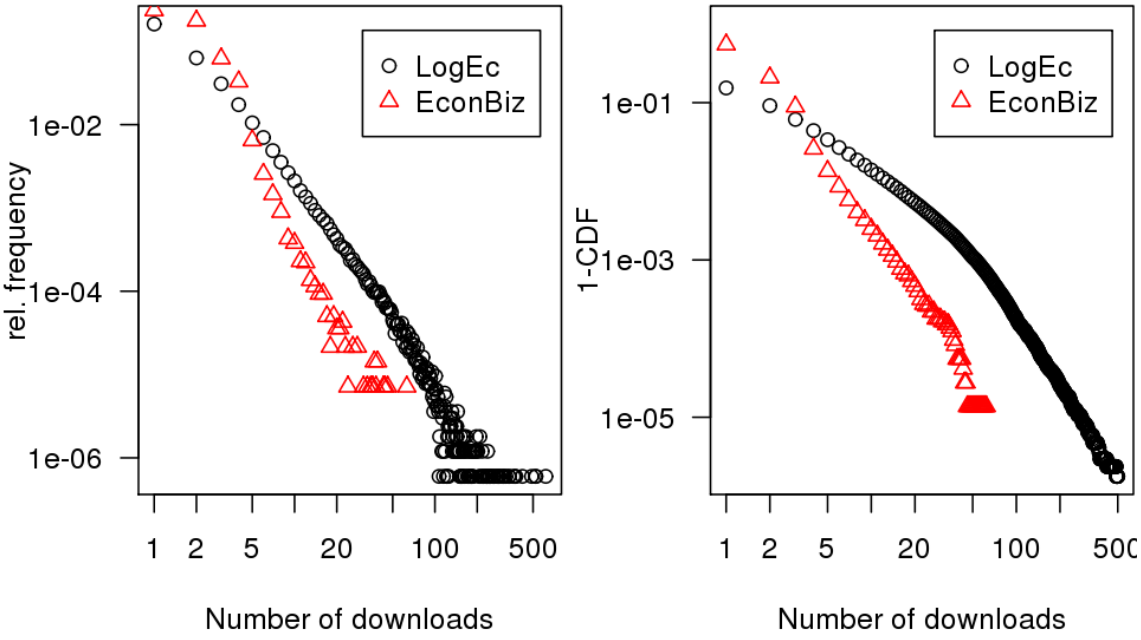

Figure 1: Number of downloads relative frequencies (left), counter cumulative distribution (right).

2.1.1 CSS score transformation

Both samples are highly left skewed and we can therefore apply the CSS method described above to construct the scales. In Table 1 we give the class boundaries and the class proportions according to the CSS method with $k_{max} = 8$ classes (documents with zero downloads have been excluded). The class proportions are quite different for the two samples. The most significant differences are due to the artifacts in the EconBiz sample for smaller number of downloads. Though, the alignment of the scales given by the class boundaries seems reasonable. The non-linear character of the alignment is clearly apparent.

Class	LogEc	Size (%)	EconBiz	Size (%)
1	3	80.415	2	45.18
2	10	14.460	3	33.95
3	23	3.622	4	11.96
4	44	1.022	5	6.32
5	72	0.333	7	1.74
6	111	0.104	10	0.61
7	168	0.031	17	0.17
8	616	0.013	63	0.08

Table 1: CSS class boundaries and class proportions for the LogEc and EconBiz usage data.

In the first version of the relevance model we will apply the CSS method as described in the previous section with a linear map of the classes to the score scale. In Figure 2 the resulting score transformations for the LogEc and the EconBiz samples are shown. Due to the differences in the

sample sizes and therefore due to the differences in the maximal number of downloads observed in the samples, the transformations are quite different.

The other plot in Figure 2 shows the resulting score distribution for the LogEc sample. This is contrasted with the untransformed score distribution, where we used the scaled number of downloads $n' = n/n_{max}$ for comparability. For the transformed scores the probability mass is distributed much more evenly and the probability mass in the end of the tail is more compressed, thereby reducing the effect of outliers. The effect for the EconBiz sample is similar (not shown).

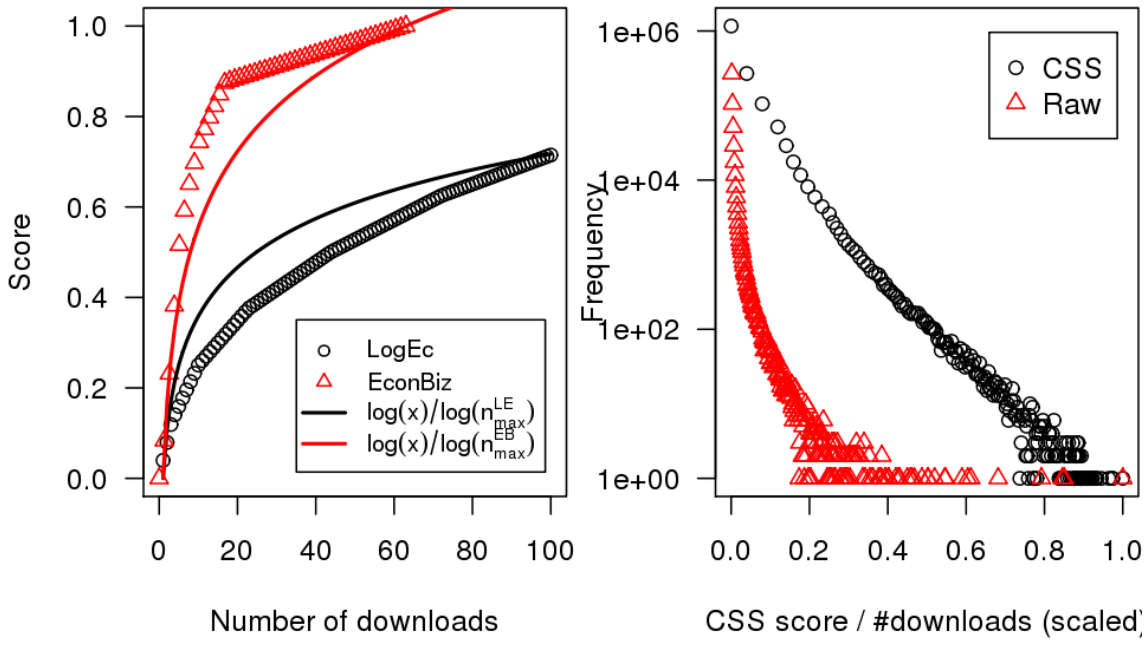

Figure 2: Transformed scores for LogEc and EconBiz. For comparision the logarithm of the number of downloads is also shown (normalized to the maximal number of downloads) (left). Comparison of the score distributions for the transformed and the raw scores for the LogEc sample (right).

2.2 Citation counts

We collected citation data from CitEc. The CitEc data covers about 630.000 documents with a total of 17.5 million references. 6.5 million of those references refer to other documents contained in RePEc. In total, the CitEc data covers about 650.000 documents that are cited at least once.

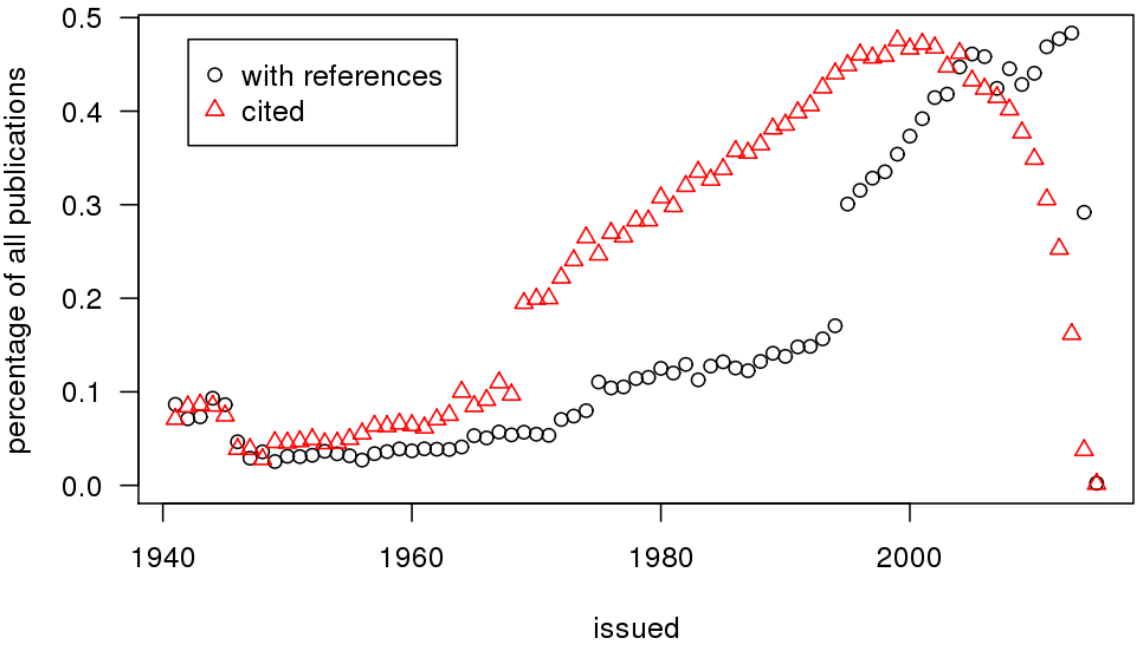

Figure 3: Coverage of publications with citation data in CitEc. Percentage of papers with reference data and percentage of papers cited at least once per year.

As is shown in Figure 3, the coverage of publications with existing reference data drops significantly around 1995. This is reflected in a drop in the percentage of cited publications around 1970. Therefore we will only consider per year distributions of citation counts back until 1970. While the coverage also drops gradually but significantly for the years before 2000, the distributions are still reasonably comparable, if one excludes the publications without citation data.

The citation count of a specific publication increases monotonically in time. Therefore, older publications should tend to have higher citation counts. Though, the percentage of the cited publications of a given age decreases (e.g. exponentially) with the age of the cited literature. This is known as citation obsolescence (e.g. Larivière, Archambault, & Gingras, 2008), which would lead to a saturation of the citation counts. On the other hand, the citation counts also depend on the growth of the literature, i.e. the citation count of a publication tends to be higher, if the number of succeeding publications is increasing. Based on a simple model assuming an exponential growth model for the number of publications (literature growth) and an exponential decay model for the citation distribution, it was shown by Egghe (1993) that the combination of both effects causes that the average citation counts rise for recent publications, reach a maximum and then decline for older publications (for typical growth and decay rates).

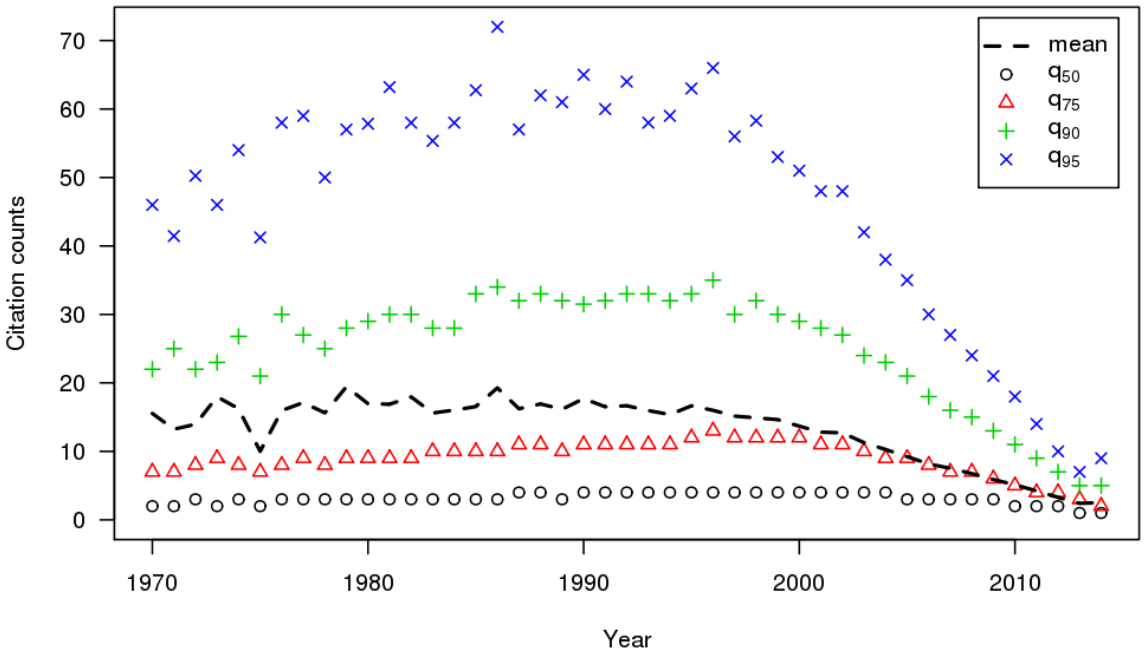

Figure 4: Comparison of the citation count distributions by the age of the publications.

Figure 4 gives a comparison of the citation count distributions by the age of the publications. For every year selected quantiles of the citation count distribution for publications issued in the given year are shown (publications without citation data excluded). All quantiles up to the 95% quantile show qualitatively the theoretically predicted behavior. Hence, citation counts are not a priori comparable between publications of different age. Though, we note that the observed behavior might also be affected by artifacts in the CitEc data and the reduced sample size for older publications.

The citation count distribution (for a specific year) follows over a large range a discrete power law distribution (or, as has been noted, equally well a log-normal or stretched exponential). In the seminal work of Price (1976) it has been shown that this can be explained by a cumulative advantage effect, also known as preferential attachment, i.e. the probability that a paper gets cited is proportional to the number of citations it had already received.

Though, the extreme counts in the tails of the distribution show a different behavior. The tails of citation distributions have been analyzed by Golosovsky & Solomon (2012). They found deviations from the power law in the tails, which could not be explained by linear preferential attachment alone. Instead they argue that the preferential attachment becomes more superlinear for older publications. Superlinear preferential attachment leads to a more concentrated accumulation of citation counts in the tails. This means that highly cited publications get even more citations than for linear preferential attachment, which leads to even longer tails. Therefore, the citation counts of highly cited publications do not show a similar decline as the quantiles shown in Figure 4.

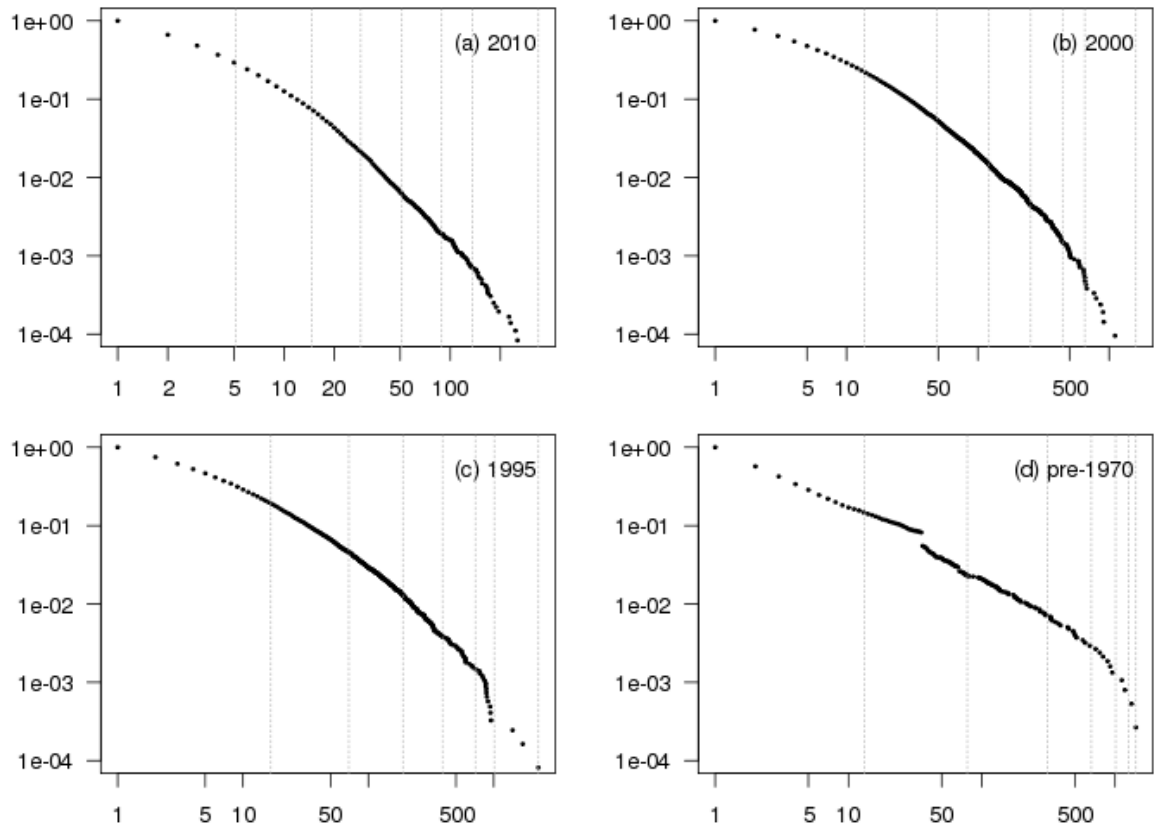

Figure 5: Counter cumulative distribution for the citations in the years 2010, 2000, 1995, and pre 1970. Dotted lines represent the corresponding CSS class boundaries.

The superlinear preferential attachment gets most apparent as an onset in the tails visible in many cumulative citation distributions (e.g. Wallace, Larivière, & Gingras, 2009). In Figure 5 we show the cumulative citation distributions for selected years for the CitEc data. The vertical lines represent the CSS class boundaries calculated for each distribution. Remarkably, the fifth CSS boundary often coincides with the beginning of the onset (e.g. for 1995 or 2000), which might simply be due to the sensitivity of the truncated mean to sudden changes in the distribution. In Golosovsky & Solomon (2012) it was further shown that the publications in the onset have very distinct characteristics from the rest of the population like an auto-correlation of the growth rate in succeeding years of almost unity, and therefore an almost deterministic growth. Consequently, they proposed that the publications in the onset are very prominent candidates for being *citation classics*.

The effectiveness of citation counts for information retrieval has been evaluated by Meij & Rijke (2007) and Zhao et al. (2014). Both studies integrated citation count data as priors into a language model. The only robust improvement with respect to mean average precision (MAP) was found by Meij & Rijke (2007) when using the (Bernoulli) maximum likelihood estimate for the citation counts for the TREC Genomics track collection, while Zhao et al. (2014) did not find an improvement for the iSearch collection. Zhao et al. (2014) also evaluated a citation induced PageRank prior, which did not result in an improvement. None of the evaluated priors did take into account the time dependence of the citation counts as discussed above.

To have a reference point we will include the citation counts for the overall citation distribution in the first version of the relevance model. The citation counts will be transformed via the CSS method.

Additionally we will calculate the per year CSS scores. For publications issued before 1970 the CSS scores will be calculated for the union of citation counts before 1970. The CSS class boundaries and class proportions are given in Table 2 for the overall distribution as well as for the selected years as in Figure 5.

k	Overall	Size(%)	2010	Size(%)	2000	Size(%)	1995	Size(%)	<1970	Size(%)
1	10	80.51	5	76.01	14	78.27	17	81.10	13	85.92
2	41	15.02	15	16.88	49	16.36	70	14.33	78	11.79
3	112	3.37	29	5.00	122	3.91	189	3.28	310	1.60
4	261	0.81	51	1.53	253	1.01	391	0.91	662	0.43
5	529	0.21	89	0.39	443	0.30	714	0.23	1009	0.16
6	971	0.06	136	0.12	652	0.10	1011	0.12	1264	0.05
7	4125	0.02	338	0.07	1599	0.05	2272	0.02	1431	0.05

Table 2: CSS class boundaries and corresponding class proportions for the overall citation distribution and for the years 2010, 2000, 1995, and pre 1970.

2.3 H-index

The h-index, proposed in Hirsch (2005), is calculated for authors registered in the RePEc Author Service[3] (RAS) based on the citation data from CitEc. It has been shown analytically in the Lotkaian framework that the h-index is tightly related with the total number of papers of an author via $h = T^{1/\alpha}$, where T is the total number of papers and α is the parameter of the Lotka distribution (Egghe & Rousseau, 2006). This has been corroborated empirically in Kelly & Jennions (2006). This indicates that the distribution of the h-index also follows a power law, which is actually the case for the CitEc data.

Hirsch assumed that the h-index should approximately increase linearly with time, and therefore proposed to use the m quotient, $m = h/n$, to compare scientists of different seniority, where n is the scientific age of a scientist, i.e. the number of years since his first publication (Hirsch, 2005). While the model of Hirsch for the publication and citation process is very simple, more realistic models corroborate the approximately linear growth (Burrell, 2007; Guns & Rousseau, 2009). Empirically, also slightly non-linear behavior has been observed (e.g. Liang, 2006).

[3] https://authors.repec.org/

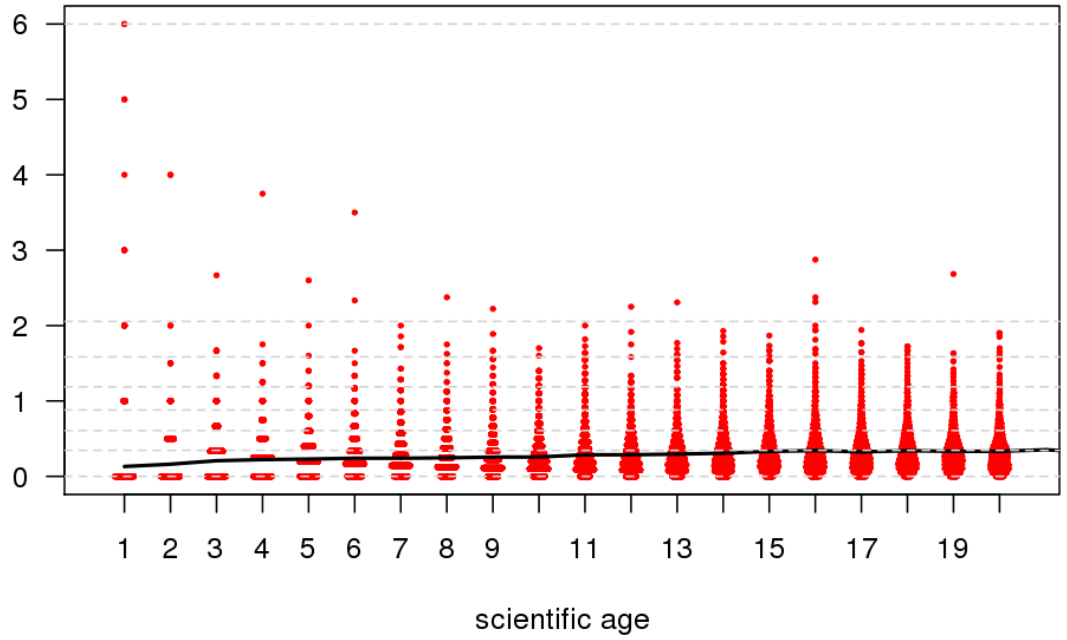

Figure 6: Distribution of the m quotient.

While the linear relationship is only an approximation, we will nevertheless use the m quotient in the first version of the relevance model to compensate or at least reduce the effect of seniority in the h-index. We note that the h-index for the CitEc data saturates for a scientific age of about 30. It is not clear if this is an artifact of the data or if this is due to some real cause, perhaps related to citation obsolescence. Though, we limit the scientific age in the calculation of the m quotient to $\mathrm{n}_{\max} = 30$, i.e. $\mathrm{n} \mapsto \min(\mathrm{n}, \mathrm{n}_{\max})$.

In Figure 6 the per age distributions of m for a scientific age up to twenty years are shown. From visual inspection we conclude that the distributions are reasonably uniform. Though, for lower levels of seniority there exist more expressed outliers, which is in part due to the discrete nature of the h-index. The mean still has a slight age dependence and therefore scientists of lower seniority will have a slight disadvantage on average. Though, one could argue that the m quotient for scientists with higher seniority is a descriptive measure for the work of the scientist, while for lower seniority it is more prescriptive and therefore less informative.

The overall m distribution is highly left skewed and its tail follows a power law. The final scores for the relevance model are determined by the CSS method, thereby again reducing the effect of the outliers. Publications with several authors will get the average score of the authors. As in the previous sections, we list the CSS class boundaries and the class proportions in Table 3.

Class	m	Size (%)
0	0.00	19.4
1	0.34	50.9
2	0.61	19.3
3	0.88	6.8
4	1.19	2.5
5	1.59	0.8
6	2.06	0.3
7	6.00	0.1

Table 3: CSS class boundaries and class proportions for the m quotient.

3 Relevance model

The relevance model will be based on the factors collected in the report[4] of work package 1. The structure of the following model description is oriented on the one of the report.

In this section these factors will be concretized and it will be defined how they are applied in the concrete use case of the EconBiz test environment[5]. It will be described from what data the factors or sub factors are constituted as well as how the data will be integrated into the relevance model. For example, for categorical factors the different levels in the data are listed, while for numerical factors it is described how the data will be transformed to utility values (e.g. by applying the CSS method; see section 1.2.3).

As mentioned in the section "Aggregation step", the overall retrieval status value will be determined by the following aggregation function:

$$\mathrm{RSV}(x) = \mathrm{score}_q \cdot \left(1 + \alpha_{qi} \cdot \sum \alpha_i \cdot v_i\right).$$

3.1 Factor Group Text Statistics

For the factor group text statistics the current ranking scheme from EconBiz will be used. The current scheme is based on a simple tf-idf field-based weighting scheme, e.g. matches in the title field are weighted higher than matches with subject indexing terms (*position of search terms*). Further, matching the complete query as a phrase is weighted higher than matching single keywords (*search term distance* and *search term order*). For queries with more than 2 or 3 words, matching sub phrases of length 2 or 3 in the query gives an additional boost. This is especially intended for boosting matches of multi-word subject indexing terms.

[4] See "Ranking: State of the Art in Web Search and Library Catalogs"
[5] See "Working Paper on Evaluation Methods within the LibRank Project", Figure 1.

3.2 Factor Group Popularity

3.2.1 Click popularity

This factor contains three sub factors: a) the **number of clicks** on a bibliographical record (views), b) clicks on (linked) tables of contents or abstracts (**further content clicks**) and c) **number of clicks on the availability button**. In EconBiz, clicks on full text links for electronic materials can be treated equally to clicks on the link resolver button for printed materials.

Although records of the document type book are clicked less often in comparison to full texts, they should not be discriminated. In relation to several theories regarding human behavior, the *Principle of Least Effort* (Zipf, 1949) explains why users are more likely to click on full text links (Case, 2009). With the particular information being only one click away, accessing it easily and instantly seems to be more important or valuable than making the effort to look at printed books on-site or borrow them in person. Though, further analysis of the usage data gathered for EconBiz is needed. Therefore, in the first version of the relevance model no distinction by type is applied.

Utility functions:

Time frame of data: last 12 months („floating window“) and last 3 months

Number of views, further content clicks, and number of clicks on the availability button will be transformed via the CSS method (see section 2.1) for both time frames.

The reason for limiting the time frame to 12 months is because one year takes all seasonal or semester caused fluctuations into consideration. Nonetheless, we will not be able to test this parameter due to the fact that we started tracking click data only about 8 months ago.

3.2.2 Dwell time

The **dwell time on a bibliographical record** is not considered as a separate factor, as the dwell time on the surrogate does not reflect the interest of a user properly. The time scale for failed information searches does not differ significantly from the dwell time for success indicating actions.

3.2.3 Usage frequency

A **record download** complies with all activities a user does with the record, that are not clicks on the availability or download or table of contents button, i.e. all external links within the record, but clicks on the buttons of the toolbar, that is the same in every record. These are: “Cite”, “E-Mail”, “Export” and “Save to memory list”.

The **number of full text downloads** cannot be measured explicitly, as one click on the download button and the actual download are treated equally. Thus, full text usage is measured by number of clicks on the availability button (see 3.2.1). Regarding **printed materials**, the number of loans is taken from ECONIS. We only consider circulated copies without renewals or extended loan periods by the same person, because these information are of little relevance in the context of usage frequency.

Utility functions:

Time frame for data: last 12 months (with "floating window") and last 3 months

Number of loans (for ECONIS data) follow a power law. While the range of the data does not span multiple magnitudes as for other power law data, the CSS method is still applicable. Therefore, for number of loans the CSS method will be applied.

3.2.4 Purchasing behavior

With purchasing behavior as an indicator for popularity, we make the general distinction between local and global demands. **On the local level**, **all copies, regardless of their location within the library are being counted**, i.e. all copies in open or closed stacks or on course reserve. Since the ZBW is not a university library, there are, for example, no course reserves or other university related stock developments. This factor is only applicable to books purchased by libraries with a rather demand-driven acquisition policy, for example, public or academic libraries that focus on their users' acquisition proposals or follow the patron-driven acquisition model (Allison, 2013). As it is part of the supra-regional literature supply system of the German Research Foundation, the German National Library for Economics (ZBW) collects all theoretical and empirical literature in economics and business studies, focusing on national customers' demands[6] and thus, represents a special case regarding collection profiles or acquisition policy.

On a global level, the editions of different formats of a work are being counted, since these data reflect the global demand (neglecting differences in the size of the editions). The number of editions for a work will be determined via the OCLC Work LOD dataset[7]. Though, the raw data gives only the number of variants of a work, i.e. not only different expressions or manifestations (in FRBR terms) are counted, but also e.g. different cataloguing variants of the same manifestation, etc.. Attempts have been made to find clusters on the manifestation as well as the content level, but the results are not available in the LOD dataset, yet (Gatenby, Greene, Oskins, & Thornburg, 2012). Nevertheless, we will use the number of variants in the OCLC Work set as approximation for this factor.

Utility functions:

Number of variants of a work follow a power law. Therefore the CSS method will be applied.

In contrary, the **number of item owning libraries** will be considered within the particular library network. The ZBW is part of the GBV (Common Library Network[8]), so the data from the GBV libraries can be implemented for this factor (for documents in ECONIS).

Utility functions:

The number of item owning libraries follows a mildly left skewed distribution. The data will therefore be transformed by the CSS method.

[6] http://www.zbw.eu/en/about-us/profile/collection-guidelines/
[7] http://www.oclc.org/data.en.html
[8] http://www.gbv.de/

3.2.5 Publisher authority

Within this popularity factor we make the distinction between reviews, books and whether journals are peer reviewed or not. Although it may be interesting to consider global numbers, the data basis for this factor will be the data available within EconBiz. Thus, we are able to focus on the publisher relevant for the particular discipline, because the ZBW collects and documents economic information, as already mentioned above concerning the factor purchasing behavior.

The original idea behind the sub factor **number of reviews** was to consider the (expert) status of reviewers but this kind of authority is based on data that will be used for measuring *reference popularity* (see 3.2.7). Thus, we will consider the number of reviews by a certain publisher as an indicator for quality. Since a review is a critical evaluation or kind of recommendation, publishers who attach value on their published books being reviewed, assumingly establish a good reputation. This means, regarding this sub factor we measure the publishers' reputations through the number of books being reviewed, for example, a publisher with 50 reviews will be weighted higher than a publisher accounted for 30 reviews.

Although, publishers' reputations are not easily detectable, we define the **number of items by a certain publisher** as a criterion. This sub factor will be based on all books published by a publisher.

In order to identify the publishers properly, their data should be standardized. Although this is not the case in EconBiz, it would only affect a very small amount of data, so that this issue can be neglected.

Utility functions:

The number of items by a certain publisher follows a power law. Therefore the data will be transformed by the CSS method.

In contrary to books and reviews, another sub factor considers the authority of a publisher by journal reputation. **Peer reviewed journals** are of higher reputation and obtain more trust in quality, than **non-peer reviewed journals** (Nicholas, 2013, p. 26)**.** Therefore, articles published in these are weighted higher than articles published in non-peer reviewed journals.

Utility functions:

Only two conditions are possible: peer reviewed or non-peer reviewed, i.e. peer reviewed journals are boosted.

Only the latest information whether a journal is peer reviewed or not will be used. This means that an article originally published in a non-peer reviewed journal by then would be wrongly weighted higher, if the journal became a peer reviewed one in the meantime. This issue can be neglected, as it is neither common practice nor traceable.

3.2.6 Ratings

The usage of rating systems is considered low in economics and no source for ratings is known. Therefore, this factor will not be integrated into the relevance model yet.

3.2.7 Reference counting

Citation counts are the most common impact measure on the item level. On the journal and author level derived measures like the impact factor or the h-index will be used instead (see *reference popularity*).

Utility functions:

As discussed in section 2.2, citations counts will be transformed via the CSS method for

a. the overall citation distribution and
b. the per year distributions, aggregated distributions for publications issued before 1970.

3.2.8 Reference popularity

This factor covers the citation impact for a journal and an author as indicators for reference popularity.

As a measure for journals the SCImago Journal Rank[9] (SJR) will be used. Only the last available SJR for a journal will be used for all articles published in the corresponding journal, i.e. older articles will not be evaluated with the SJR from the corresponding year of publication. This is based on the assumption that from a user perspective only one quality estimate for a journal exists (i.e. the current one) and not a full history of quality estimates.

The scientific impact of an **author** will be measured on the basis of his or her m quotient (see section 2.3). The m quotient is based on the citation counts from CitEc. While the citation counts from CitEc are biased due to the selection of works integrated in RePEc, the author data from RAS can be regarded as one of the most complete sources for disambiguated authors in economics joined with their corresponding works including reference data.

Utility functions:

The m quotient will be transformed by the CSS method (see section 2.3).

3.3 Factor Group Freshness

3.3.1 Publication date

The **publication dates of reviews, books and articles** can be seen as a common indicator for freshness, because without a particular interest of publishers, researchers, authors, or other parties, works would not be published at all. Thus, this ranking factor is of high significance, as well as it is generally used for presenting search results in different information systems, not only in the library context.

[9] http://www.scimagojr.com/

Journal **articles** are often being published online first and the printed version (parallel output) might be available at a much later point in time, for example, one year later. In addition, publication dates of online articles usually are of a granular format, i.e. year, month and day; whereas publication dates of **printed books** commonly consist of the year. Therefore, the question rises, whether to make the distinction between "printed" and "electronic" materials. If a work is made available online first, then this date should count as publication date, but since the parallel output(s) of one particular work is currently catalogued as a separate library record[10], a joint data integration of publication dates of those records cannot be made.

A distinction between the different **date formats** does not have to be made, because to ensure comparability within the assessment model, the date format used will only contain year dates.

Documents of the **type** "article" and "book" could be treated differently regarding their time constants, as the perception of what current literature covers exactly is different as well. In EconBiz the document type "book" is also used for working papers, which are of high importance in economics (Zimmermann, 2013).

The distinction could be made on content level, i.e. research area, or on methodology level, i.e. theoretical or empirical level. We can assume that economists tend to judge current literature not to be more than 2 to 5 years old, i.e. articles published in the last 2 years and books not older than 5 years. Working papers can be assumed to be even more important as this type contains most current research results, due to the publishing process in general.[11]

In order to avoid highly cited literature published decades ago (classics) being ignored, this ranking factor must be applied in combination with popularity factors, for example *reference popularity* or *usage frequency*.

Utility functions:

The idea is to test this freshness factor in the first evaluation run with the currently used boost score, but without the distinction between document types, and to set **limits on the time constant** for publication dates at 2 and 5 years. In later runs, the limits should be altered and a distinction between the document types and, perhaps even content, should be evaluated.

3.3.2 Accession date

The day the library acquired (or licensed) a work is the accession date. It is assumed, that the time between acquiring a work and the day it is made available after processing, can be neglected, as that period would not be an indicator for freshness. Instead, the circumstance that old books covering one particular topic and suddenly are being acquired by the library, suggest a renewed interest which, again, indicates freshness.

[10] With regard to the cataloging rules RAK-WB.

[11] For sharing current research results it is commonly known that articles in general are more appropriate as a publishing format than, for example, books, regarding the time span of the entire process of writing and publishing.

As mentioned above with regard to the factor purchasing behavior, this factor is also only applicable to books purchased by libraries with a rather demand-driven acquisition policy. Therefore, the factor accession date will not be considered.

3.4 Factor Group Locality & Availability

3.4.1 Physical location

The physical location of user and item is closely connected with the factor availability (see 4.2), as it is a prerequisite for defining its utility functions. Regarding on the **user's location**, we can refer to the user models[12]: A user is located within the library rooms or outside of the library rooms (e.g. at home), regardless of the network access or mobile accessibility.

Concerning the physical location of the library item, we have to differ between printed and electronic materials: We assume **printed materials** to be located in the library rooms, because they have been purchased by the library, i.e. are in stock. If they have been purchased, but are not located in the library rooms, the factor availability applies.

Electronic materials can be viewed under two aspects: 1) They are purchased or licensed by the library; 2) they have not been purchased or licensed by the library. A third viewpoint would concern the physical location (or network accessibility[13]) of the user in conjunction with *availability of the webpage / item*, but this is described in the following section, as well as the utility functions.

3.4.2 Availability of the webpage / item

The availability of an item in the library context is dependent on the physical location of the user. To clarify all possible conditions for availability, the following table gives an overview on reasons for an item being available or not:

Status	Printed materials	Electronic materials
Available	Purchased (inventory evidence); Currently not circulating	Licensed or purchased; Free
Not available	Not purchased; Currently circulating	Not licensed, nor purchased; Not free

Table 4: Definition of availability in the library context

As mentioned above, we consider two user models, but this factor demands a third one: a user is located anywhere but the library (e.g. at home) and at the same time is located within the library network, she or he would be able to access electronically available materials as if her or his location would be on-site.

[12] See "Working Paper on Evaluation Methods within the LibRank Project", section 3.1.

[13] In the academic world, if a user is not located in the library but in the office on campus, he or she would still be within the university network which allows accessing electronic library items, regardless of the physical location. A connection to the university network is also possible via the use of VPN (Virtual Private Network).

A user within the library network will be treated the same way as a user physically on-site. Though, the service of a VPN is currently not offered to ZWB users. Thus, we only consider the described two users models, i.e. users are in the ZBW at the location in Hamburg or Kiel or the users are not on-site (the information can be obtained via the IP address).

Utility functions:

Every available item (printed or electronic) gets a boost and every non-available item will not get boosted with one exception: If a printed item is not available because it is circulating, it will be boosted a little as inventory evidence. This means, whether the user is on-site or not, does not matter, as we cannot know if a user may be at home but plans to visit the library for looking at selected printed books in person.

3.5 Factor Group Content Properties

3.5.1 Additional information

As a sub factor, the **bibliographic record** must be complete, i.e. the lack of metadata may be seen as an indicator for lacking reliability.

Currently, there are a larger number of duplicate records in EconBiz due to the integration of different, heterogeneous data sources. In case of a work represented by two surrogates, the one record containing subject terms is preferred compared to the one without subject terms. By this means, the requirement of boosting only complete records is met.

Another sub factor covers **abstracts, tables of contents and reviews** as additional information that may indicate relevance (in terms of added value). A distinction has to be made between abstracts and tables of contents compared to reviews because the latter item type is represented by a separate record whereas abstracts and tables of contents are part of the metadata (or external links) of one particular record.

Utility functions:

There are only two possible conditions to be considered: the information is available or not. This means, if a record contains an abstract or table of contents it will be boosted, a record linked to a review will be boosted, and if there are abstract or table of contents and also a review linked to a record, it will get an additive boost. A problem may be, that only books could have both abstract and table of contents as part of its metadata, but nevertheless, articles must not be discriminated in this matter.

The availability of underlying **research data** will not be considered as a factor to be tested, since research data currently are considered to be in an early development phase in economics. As integrated part of metadata in other library information systems, this factor might be highly relevant and should be tested with the necessary test environment.

3.5.2 File format

This factor will not be integrated into the relevance model, since no attempt to gather corresponding data will be made (see Functional specifications).

3.5.3 Language

A prerequisite for testing this factor is that the desired language of the user is known. For EconBiz users the language setting could be found out due to the browser setting or the language(s) selected in the advanced search interface. In general, users prefer the simple search interface, so that such data cannot be taken into consideration. Furthermore, the language setting with regard to relevance assessments cannot be evaluated systematically within this project.

3.6 User background

Due to the lack of data, this factor group consisting of the factors *user group* and *usage data* will not be integrated into the relevance model (see Functional specifications).

4 References

Allison, D. A. K. (2013). *The patron-driven library: a practical guide for managing collections and services in the digital age*. Oxford: Chandos Pub.

Bouyssou, D. (2000). Evaluation and decision models : a critical perspective.

Burrell, Q. (2007). Hirsch's h-index: A stochastic model. *Journal of Informetrics*.

Case, D. O. (2009). Principle of Least Effort. In K. E. Fisher, S. Erdelez, & L. (E. F. . McKechnie (Eds.), *Theories of Information Behavior* (pp. 289–292). Medford, New Jersey: Information Today.

Craswell, N., & Robertson, S. (2005). Relevance weighting for query independent evidence. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 416–423).

Da Costa Pereira, C., Dragoni, M., & Pasi, G. (2012). Multidimensional relevance: Prioritized aggregation in a personalized Information Retrieval setting. *Information Processing & Management*, *48*(2), 340–357. doi:10.1016/j.ipm.2011.07.001

De Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. doi:10.1002/asi.4630270505

Egghe, L. (1993). On the influence of growth on obsolescence. *Scientometrics*, *27*(2), 195–214. doi:10.1007/BF02016550

Egghe, L. (2010). Characteristic scores and scales in a Lotkaian framework. *Scientometrics*, *83*(2), 455–462. doi:10.1007/s11192-009-0009-y

Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, *69*(1), 121–129. doi:10.1007/s11192-006-0143-8

Eickhoff, C., de Vries, A. P., & Collins-Thompson, K. (2013). Copulas for information retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13* (p. 663). doi:10.1145/2484028.2484066

Farah, M., & Vanderpooten, D. (2006). A Multiple Criteria Approach for Information Retrieval. In *SPIRE, String Processing and Information Retrieval* (pp. 242–254). Springer Science + Business Media. doi:10.1007/11880561_20

Gatenby, J., Greene, R. O., Oskins, W. M., & Thornburg, G. (2012). GLIMIR: Manifestation and Content Clustering within WorldCat. *Code4Lib Journal*, (17).

Glänzel, W. (2007). Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, *1*(1), 92–102. doi:10.1016/j.joi.2006.10.001

Glanzel, W., & Schubert, a. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Studies in International Education*, *14*(2), 123–127. doi:10.1177/102831538801400208

Golosovsky, M., & Solomon, S. (2012). Runaway events dominate the heavy tail of citation distributions. *European Physical Journal: Special Topics*, *205*, 303–311. doi:10.1140/epjst/e2012-01576-4

Guns, R., & Rousseau, R. (2009). Simulating growth of the h-index. *Journal of the American Society for Information Science and Technology*, *60*, 410–417. doi:10.1002/asi.20973

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 102, pp. 16569–16572). doi:10.1073/pnas.0507655102

Kelly, C. D., & Jennions, M. D. (2006). The h index and career assessment by numbers. *Trends in Ecology and Evolution*, *21*, 167–170. doi:10.1016/j.tree.2006.01.005

Larivière, V., Archambault, É., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900-2004). *Journal of the American Society for Information Science and Technology*, *59*(2), 288–296. doi:10.1002/asi.20744

Liang, L. (2006). H-index sequence and h-index matrix: Constructions and applications. *Scientometrics*, *69*(1), 153–159. doi:10.1007/s11192-006-0145-6

Meij, E., & Rijke, M. De. (2007). Using prior information derived from citations in literature search. In *RIAO '07 Large Scale Semantic Access to Content (Text, Image, Video, and Sound) Proceedings* (pp. 665–670). Centre de Hautes Etudes Internationales d'Informatique Documentaire.

Mitzenmacher, M. (2004). A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, *1*(2), 226–251. doi:10.1080/15427951.2004.10129088

Moulahi, B., Tamine, L., & Yahia, S. Ben. (2014). Learning to Match for Multi-criteria Document Relevance, 9. arXiv preprint arXiv:1409.6512.

Nicholas, D. (2013). Trust and authority in scholarly communications. In *VI Encontro Ibérico EDICIC, Porto (Portugal), 4-6 November* (pp. 19–32). Faculdade de Letras da Universidade do Porto - CETAC.MEDIA.

Nottelmann, H., & Fuhr, N. (2003). From Retrieval Status Values to Probabilities of Relevance for Advanced IR Applications. *Information Retrieval*, *6*(3-4), 363–388. doi:10.1023/A:1026080230789

Richardson, M., Brill, E., & Prakash, A. (2006). Beyond PageRank. In *Proceedings of the 15th international conference on World Wide Web* (pp. 707–715). Association for Computing Machinery. doi:10.1145/1135777.1135881

Robertson, S. E., & Walker, S. (1994). Some for Simple Effective Approximations to the 2 – Poisson Model Probabilistic Weighted Retrieval The. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 232–241). doi:10.1016/0306-4573(94)00047-7

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, *15*(3), 234–281. doi:10.1016/0022-2496(77)90033-5

Silva, T. P. C., de Moura, E. S., Cavalcanti, J. M. B., da Silva, A. S., de Carvalho, M. G., & Gonçalves, M. A. (2009). An evolutionary approach for combining different sources of evidence in search engines. *Information Systems*, *34*(2), 276–289. doi:10.1016/j.is.2008.07.003

Wallace, M. L., Larivière, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, *3*(4), 296–303. doi:10.1016/j.joi.2009.03.010

Wang, P., & Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science*, *49*(2), 115–133. doi:10.1002/(SICI)1097-4571(1998)49:2<115::AID-ASI3>3.0.CO;2-1

Yager, R. R. (1988). Ordered Weighted Averaging Aggregation Operators in Multicriteria Decisionmaking. *IEEE Transactions on Systems, Man and Cybernetics*, *18*(1), 183–190. doi:10.1109/21.87068

Zhao, H., & Hu, X. (2014). Language Model Document Priors based on Citation and Co-citation Analysis. In *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval* (pp. 29–36).

Zimmermann, C. (2013). Academic Rankings with RePEc. *Econometrics*, *1*(3), 249–280. doi:10.3390/econometrics1030249

Zipf, G. K. (1949). *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge, Mass.: Addison-Wesley Press.