Copulas for Information Retrieval

Carsten Eickhoff
Delft University of Technology
Delft, The Netherlands
c.eickhoff@acm.org

Arjen P. de Vries
CWI Amsterdam
Amsterdam, The Netherlands
arjen@acm.org

Kevyn Collins-Thompson
Microsoft Research
Redmond, WA, USA
kevynct@microsoft.com

ABSTRACT

In many domains of information retrieval, system estimates of document relevance are based on multidimensional quality criteria that have to be accommodated in a unidimensional result ranking. Current solutions to this challenge are often inconsistent with the formal probabilistic framework in which constituent scores were estimated, or use sophisticated learning methods that make it difficult for humans to understand the origin of the final ranking. To address these issues, we introduce the use of *copulas*, a powerful statistical framework for modeling complex multi-dimensional dependencies, to information retrieval tasks. We provide a formal background to copulas and demonstrate their effectiveness on standard IR tasks such as combining multidimensional relevance estimates and fusion of results from multiple search engines. We introduce copula-based versions of standard relevance estimators and fusion methods and show that these lead to significant performance improvements on several tasks, as evaluated on large-scale standard corpora, compared to their non-copula counterparts. We also investigate criteria for understanding the likely effect of using copula models in a given retrieval scenario.

Categories and Subject Descriptors

Information Systems [**Information Retrieval**]: Retrieval models

Keywords

Relevance models; Multivariate relevance; Ranking; Probabilistic framework; Data fusion.

1. INTRODUCTION

In response to user queries, today's search systems typically return lists of documents ranked by system estimates of relevance. In traditional IR retrieval models, each document's relevance towards the query is expressed as term overlap between query and document [42]. Early on, researchers began exploring alternative, non-topical document quality criteria such as document recency, credibility or monetary cost. More recently, through a combination of improved algorithms and greatly increased data scale, significant gains in ranking quality and user satisfaction based on employing non-topical factors such as textual complexity [12] or suitability for children [17] have begun influencing the ranking process. Given a scenario such as child-friendly information search, non-topical quality criteria can clearly have a strong influence on usefulness of a document for a specific user. A perfectly relevant document that is not understandable due its complex sentence structure or excessive use of jargon will have significantly diminished user relevance.

Beyond the value of individual relevance factors, there can be complex, non-linear *dependencies* between relevance factors. For example, relevance criteria such as topicality and credibility might appear independent for some document subsets, but extreme values in one dimension may influence the other in a way that is not easily captured by state-of-the-art approaches. As a concrete example, take TREC 2010's faceted blog distillation task [32], that aims at retrieving topically relevant non-factual blog feeds. Here, the relevance space has two dimensions: topicality and subjectivity. Figure 1 shows the distribution of relevance scores for Topic 1171, "*mysql*", across these two relevance dimensions. We can note an apparent correlation in the lower left part of the graph that weakens as scores increase. To underline this, we computed Pearson's ρ between the two dimensions for the lower score third ($\rho = 0.37$), the upper region ($\rho = -0.4$), as well as the overall distribution ($\rho = 0.18$). Apparently, the dependency structure of the joint distribution of relevance, in this case, is not easily described by a linear model. Consequently, we can expect dissatisfying performance of linear combination models. And, indeed, when inspecting the performance of a linear combination model with empirically learned mixture parameters λ, Topic 1171 receives an average precision of only 0.14, well below the method's average across all topics of 0.25. In the course of this work, we will discuss practical means of addressing cases like the present one and will finally revisit this example to demonstrate the effect of our proposed method.

While the machine learning, information retrieval, data mining and natural language processing communities have significant expertise in estimating topical document relevance and additional criteria in isolation, the commonly applied combination schemes have tended to be *ad hoc* and ignore the problem of modeling complex, multi-dimension dependencies. In practice, they follow statically weighted lin-

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

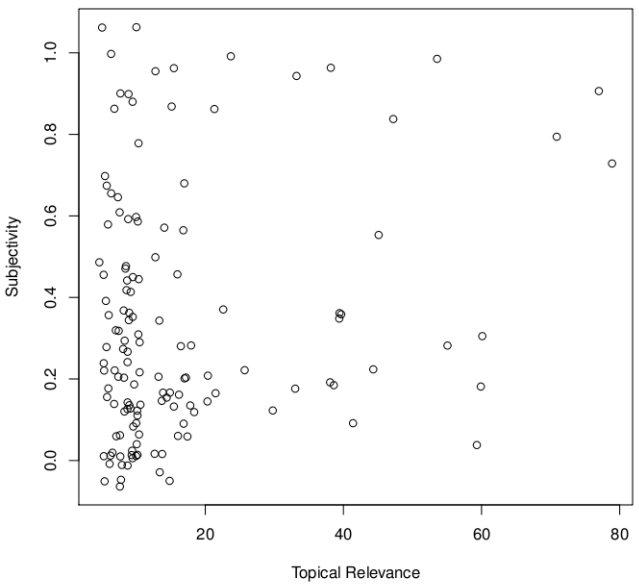

Figure 1: Distribution of bivariate relevance scores for TREC 2010 Blog Track Topic 1171, "mysql".

ear combinations with empirically determined mixture parameters [42] or deploy sophisticated learning to rank techniques that tend to offer only limited insight to humans about why they were weighted highly for relevance. Ideally, we would demand realistic, yet formally-grounded combination schemes that can lead to results that are both effective and with human-interpretable justification.

In a different context, the field of quantitative risk management has devised *copulas*, a flexible, varied class of probability density functions that are designed to capture rich, non-linear dependencies efficiently in multi-dimensional distributions. Copulas work by decoupling the marginal distributions of the data from the underlying dependency structure of the joint distribution. In particular, copulas can account for so-called tail dependencies, i.e., dependencies that play up at the extreme values of the interacting distributions. As an example, let us consider two commodities traded on the stock market, such as rare earth metals and pork bellies. The two commodities are sufficiently different to make the related market segments quasi-independent. However, extreme market situations have been shown to cause investor panics that reach across otherwise independent segments and cause previously unseen interrelationships [9].

This work makes three contributions to the state of the art in relevance modelling. **(1)** We give a detailed introduction to the formal framework of copulas and describe how to estimate them from empirical data. **(2)** Based on a number of sizeable standard data sets such as the Blogs08 collection [32], we demonstrate the merit of using copulas for multivariate relevance estimation. **(3)** In a related effort, we address the task of score fusion based on historic submissions to the TREC *ad hoc* task.

The remainder of this paper is structured as follows: Section 2 gives a historic overview of IR relevance frameworks, prior work on multidimensional relevance models, score fusion approaches, as well as, examples of copula applications from different fields. Section 3 formally introduces the theoretical foundation of copulas and details key techniques in their application. In Sections 4 and 5 we demonstrate their merit at the tasks of estimating multidimensional relevance scores as well as fusing prior TREC runs. Section 6 further discusses the experimental results and aims at identifying those domains of IR for which copulas are most promising. Section 7 concludes the paper with a concise summary of our findings.

2. RELATED WORK

Over the past decades, a wide range of partially overlapping relevance frameworks have been proposed, a few prominent examples include [44, 22, 34, 8]. They unanimously consider relevance as a complex, potentially multidimensional concept that may be composed from a number of constituents. In the further course of this section, we will focus on the practical implementation of formal relevance estimation schemes employed in information retrieval and related disciplines. Schamber *et al.* [45] radically revised the definition of relevance, causing a growing interest in probabilistic relevance modelling in the research community. First openly applied at the third TREC competition, the BM25 retrieval model [43] represents a performance landmark that is still valid today (with slight variations such as the 2004 integration of multiple weighted fields [42]). In 1996, Persin *et al.* [38] introduced the idea of retrieval result lists ranked by their probability of relevance, as an alternative to the previously dominant binary retrieval scenario. Two years later, Ponte and Croft proposed the use of language modelling techniques to determine topical relevance [39]. One of the first notions of non-topical relevance was expressed in Kleinberg's work on hubs and authorities [26] in which the author introduces two document-specific relevance notions independent of the query. Lavrenko and Croft [28, 29] pursued a line of work on dedicated relevance models.

While the formal combination of several individual relevance facets in one model has not been extensively studied, there has been an interesting thread of research on score fusion. The task is to combine the result rankings of multiple independent retrieval systems in order to compensate for local inaccuracies of single engines. Early approaches to the task were based on evidence aggregation in the form of products and sums of scores across individual systems [19]. The fused ranking is based on the absolute value of the cross-system aggregates. Vogt *et al.* [51] first introduced linear interpolation of multiple model rankings for system fusion. Aslam and Montague [5] proposed a probabilistic rank-based method for direct combination of multiple engines. Later on, they devised a similar method based on a majority voting scheme between various retrieval systems [35]. [36] proposed a score normalization scheme that is more robust to outliers in the distribution of relevance than the previously used min/max technique. There has been an extensive body of work on estimating the distribution of relevance scores for document ranking. Recent examples include the work by Arampatzis and Stephenson [4], Kanoulas *et al.* [25], and, Cummins [14]. Manmatha et al. [33] estimated a search engine's score distribution as a mixture of normal and exponential distributions, for relevant and non-relevant documents respectively. They

used the resulting distributions for score fusion across multiple engines, but did not attempt to model dependencies in the joint score distribution, instead treating the scores as independent and averaging probabilities, or discarding 'bad' engines altogether.

In 2002, Wu and Crestani [52] introduced the first of what would become a group of fusion approaches that define an explicit weighting scheme under which the original result lists are combined. [7] and [15] employ various quality notions such as the degree to which a document satisfies a given relevance criterion to dynamically adapt the weighting scheme to the underlying distribution of relevance. In 2005, Craswell *et al.* investigated relevance model combination by linearly combining constituent scores in the log domain [13]. Tsikrika and Lalmas applied Dempster-Shafer theory for the aggregation of independent relevance criteria in web retrieval in the form of belief functions [49]. Gerani *et al.* [21] propose non-linear score transformations prior to the standard weighted linear combination step. Their solid results demonstrate the need for models whose capabilities go beyond linear dependency structures between relevance dimensions.

In recent years, the variety of IR applications has become significantly more diverse. As a consequence, universal relevance models have become less viable in many areas. Tasks such as legal IR, expert finding, opinion detection or the retrieval of very short documents (e.g., tweets) have brought forward strongly customised relevance models tailored towards satisfying a given task (e.g., [24, 6]). Especially for the retrieval of structured (XML) documents, score combination schemes are of central importance to combine evidence across multiple structural fields within a document. Despite the numerous potential issues pointed out by Robertson *et al.* [42], most state-of-the-art approaches to XML retrieval rely on linear models [31]. An advance towards the formal combination of several independent relevance criteria in the form of prior probabilities for language models has been made by Kraaij *et al.* [27] for the task of entry page search. To date, however, most universally applicable relevance models still rely on pure linear combinations of relevance criteria that disregard the underlying data distribution or potential dependencies between the considered dimensions.

Learning to rank (L2R) has been established as an alternative approach for signal combination. The aim is to apply machine learning methods to either directly infer a document ranking or a ranking function from a wide range of features, potentially including the previously-discussed relevance criteria [10, 40, 30]. The downside of this approach is that the resulting models tend to yield only limited insight for humans. The classic approach of developing a unifying formal retrieval model would in our view provide better means to increase not just overall performance, but also our qualitative understanding of the problem domain.

By introducing copulas for information retrieval, this work proposes a way for closing the gap between linear combinations (that break with the probabilistic framework in which the constituent scores were estimated) and non-linear machine-learned models (that offer only limited insight to scientists and users).

Copulas have been traditionally applied for risk analyses in portfolio management [18] as well as derivatives pricing [9] in quantitative finance. Recently, however, there are several successful examples from unrelated disciplines. Renard *et al.* estimate water flow behaviour based on Gaussian copulas [41]. Onken *et al.* apply copulas for spike count analysis in neuroscience [37]. In meteorology, copulas have been used to combine very high-dimensional observations for the task of climate process modelling [47]. To the best of our knowledge, there has been no prior application of the copula framework to information retrieval problems.

3. COPULAS

At this point, we will give a brief introduction of the general theoretical framework of copulas, before applying them to various IR tasks in subsequent sections. For a more comprehensive overview, please refer to [46] for more detail and pointers to further reading.

The term copula was first introduced by Sklar [48] to describe multivariate *cumulative distribution functions* (*cdfs*) that allow for a formal decoupling of observations from dependency structures. Formally, given

$$X = (x_1, x_2, \ldots, x_k)$$

a k-dimensional random vector with continuous margins

$$F_k(x) = \mathbb{P}[X_k \leq x]$$

we can map our observations to the unit cube $[0,1]^k$ as

$$U = (u_1, u_2, \ldots u_k) = (F_1(x_1), F_2(x_2), \ldots F_k(x_k)).$$

This is where our copulas come into play. A k-dimensional copula C describes the joint cumulative distribution function of random vector U with uniform margins.

$$C : [0,1]^k \rightarrow [0,1]$$

This approach has two obvious practical benefits: (1) Separating marginals and dependency structure allows for more straightforward estimation or approximation of each component in isolation. (2) An explicit model of dependency is scale-invariant. The copula describes a reference case of dependency on the unit cube $[0,1]^k$ that can be applied to arbitrary random vectors without further adjustment.

A number of key properties make copulas an appealing theoretical framework for a wide number of applications, so we summarize those now.

- Like all cdfs, a copula $C(u_1, u_2, \ldots, u_k)$ is increasing in each component u_i
- A marginal component u_i can be isolated by setting all remaining components to 1:

$$C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$$

- If a single component u_i in U is zero, the entire copula is zero:

$$C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_k) = 0$$

- Most importantly, we can assume general applicability of the copula framework, since, as a consequence of Sklar's Theorem [48], for each k-dimensional cdf F and all x_i in $[-\infty, \infty]$ and $1 \leq i \leq k$, there exists a copula C with

$$F(x_1, \ldots, x_k) = C(F_1(x_1), \ldots, F_k(x_k))$$

3.1 Extreme conditions

Before applying the copula framework to problems in information retrieval, let us visit a number of extreme conditions of dependency that frequently occur in IR scenarios. **(1) Independence** of observations is a frequently assumed simplification in IR theory that leads to convenient (if naïve) probabilistic models. In the copula framework, independence of events can be captured by the so-called *independence copula* C_{indep}:

$$C_{indep}(U) = \exp(-\sum_{i=1}^{k} -\log u_i)$$

which is equivalent to the product across all constituent probabilities in U. **(2) Co-monotonicity** describes the case of perfect positive correlation between observations u:

$$C_{coMono}(U) = \min\{u_1, \ldots, u_k\}$$

(3) counter-monotonicity of observations is given in the opposite case of perfect negative correlation:

$$C_{counterMono}(U) = \max\{\sum_{i=1}^{k} u_i + 1 - k, 0\}$$

Consequently, each copula lies within the so-called Fréchet-Höffding bounds [23]:

$$C_{counterMono}(U) \leq C(U) \leq C_{coMono}(U)$$

3.2 Copula families

After having covered the foundations of copula theory let us inspect some concrete examples of copulas that will be used in the course of this work. Three general families of standard copulas have been proposed in the literature, whose corresponding equations are given right after their introduction in this paragraph: **(1) Elliptical copulas** are directly derived from known distributions and are based on standard distribution functions such as the Gaussian distribution or Student's t distribution. Equation 1 shows the Gaussian copula that requires the observed covariance matrix $\Sigma \in \mathbb{R}^{k \times k}$ as a parameter. Φ denotes the cdf of a standard normal distribution and Φ^{-1} its inverse. **(2) Archimedean copulas** are popular as they can be explicitly stated (note that due to their distribution dependency that is not the case for elliptical copulas) and typically depend on only a single degree of freedom. The parameter θ expresses the strength of dependency in the model. Equation 2 shows the Clayton copula whose θ-range is $[-1, \infty) \backslash \{0\}$. $\theta = -1$ represents counter-monotonicity, $\theta \to 0$ gives the independence copula and $\theta \to \infty$ approaches co-monotonicity. Finally, **(3) Extreme value copulas** are robust in cases of extreme observations. The Gumbel copula (Equation 3) has a parameter space of θ in $[1, \infty)$. For $\theta = 1$ we obtain the independence copula, and, for $\theta \to \infty$ we approach co-monotonicity.

$$C_{Gaussian}(U) = \Phi_{\Sigma}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_k)) \quad (1)$$

$$C_{Clayton}(U) = (1 + \theta(\sum_{i=1}^{k} \frac{1}{\theta}(u_i^{-\theta} - 1)))^{\frac{-1}{\theta}} \quad (2)$$

$$C_{Gumbel}(U) = \exp(-(\sum_{i=1}^{k} (-\log(u_i))^{\theta})^{\frac{1}{\theta}}) \quad (3)$$

Figure 3.2 shows contour plots of a number of bivariate standard copulas. The concrete choice of copula family and instantiation has been frequently reported to depend on the application domain [46]. If no prior knowledge about the dependency structure, e.g., prevalence of asymptotic or tail dependencies, is available, practitioners often resort to goodness-of-fit tests or measures of tail dependency in order to choose an appropriate model. We will describe the use of these techniques in the subsequent sections when applying copulas for information retrieval problems.

3.3 Fitting copulas to observations

In the case of elliptical copulas, the fitting process is limited to calculating means and covariance matrices from the available observations. Here, the only degree of freedom is the concrete choice of distribution function (e.g., Gaussian vs. Student) that best approximates the original distribution that generated the observations. In the non-elliptical case, the task is to determine optimal settings of θ. Commonly, this is achieved by means of maximum likelihood estimates based on the available observations. This is also the approach chosen in this work. It should be noted that there are methods for direct empirical estimations of entire copula functions. The interested reader can find a good overview by Charpentier *et al.* [11] as a starting point for this line of research, the inclusion of which would however go beyond the scope of this initial exploration of copulas for information retrieval.

4. RELEVANCE ESTIMATION

In the previous section, we described the theoretical foundations of copulas including concrete ways of computing $C(U)$ from multivariate observations U. We now detail their application for relevance estimation in information retrieval. First, we separately estimate the probability of relevance $P_{rel}^{(k)}(d)$ and non-relevance $P_{non}^{(k)}(d)$ for a document d, under each of the k criteria (dimensions) – for example, topicality, recency, readability, etc. Next, we assume random observations U_{rel} and U_{non} to derive from these distributions and base two distinct copulas, C_{rel} and C_{non} on them.

Recall that these copulas should capture the dependencies between relevance criteria, in either the relevant (C_{rel}) or the non-relevant (C_{non}) documents retrieved. Since it is difficult to predict where these dependencies have the most effect, it is natural to consider three different general approaches of combining multivariate observation scores U into a single probability of relevance that can be used for resource ranking. **(1)** $CPOS(U_{rel})$ multiplies the independent likelihood of observing U_{rel} with the relevance copula C_{rel}, capturing only dependencies between the likelihoods of relevance. **(2)** $CNEG(U_{rel}, U_{non})$ normalizes the probability of relevance by the non-relevance copula $C_{non}(U_{non})$, capturing only the dependencies between the likelihoods of non-relevance. **(3)** $CODDS(U_{rel}, U_{non})$, finally, multiplies the probability of relevance by the ratio of the two copulas, modelling simultaneously the dependencies between both previous notions.

$$CPOS(U_{rel}) = C_{rel}(U_{rel}) \prod_{i=1}^{k} u_{rel,i}$$

$$CNEG(U_{rel}, U_{non}) = \frac{\prod_{i=1}^{k} u_{rel,i}}{C_{non}(U_{non})}$$

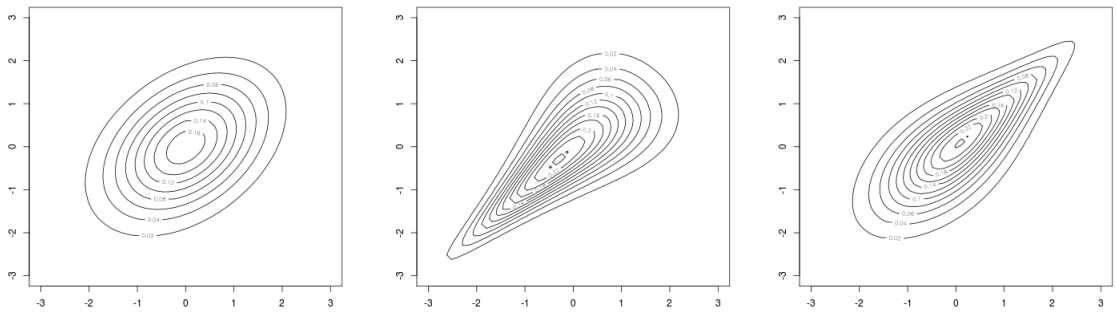

Figure 2: Examples of bivariate copula contour plots. (a) Gaussian copula, (b) Clayton copula with $\theta = 2.0$, (c) Gumbel copula with $\theta = 2.0$.

$$CODDS(U_{rel}, U_{non}) = \frac{C_{rel}(U_{rel})}{C_{non}(U_{non})} \prod_{i=1}^{k} u_{rel,i}$$

As performance baselines, we will compare to three popular combination methods from the literature: **(1)** $SUM(U_{rel})$ sums up the relevance scores across all criteria k and uses the sum as the final ranking criterion [19]. **(2)** $PROD(U_{rel})$ builds the product across all constituents [19]. Probabilistically, this combination scheme assumes independence across all criteria and can be expected to be too naïve in some settings where dependence is given. **(3)** Weighted linear combinations $LIN_{\Lambda}(U_{rel})$ build a weighted sum of constituents $u_{rel,i}$ with mixture parameters λ_i optimized by means of a parameter sweep with step size 0.1 [51]. It should be noted that all optimizations and parameter estimations, both for the baselines as well as for the copula models are conducted on designated training sets that do not overlap with the final test sets. We relied on the original training portion of the respective corpora. In the case that the original corpus did not specify a dedicated training set, we used a stratified 90%/10% split.

$$SUM(U_{rel}) = \sum_{i=1}^{k} u_{rel,i}$$

$$PROD(U_{rel}) = \prod_{i=1}^{k} u_{rel,i}$$

$$LIN_{\Lambda}(U_{rel}) = \sum_{i=1}^{k} \lambda_i u_{rel,i}$$

Based on three different standard datasets and tasks, we will highlight the merit of using copulas over the traditional approaches. Each of the settings specifies 2 individual relevance criteria ($k = 2$) which are crucial for user satisfaction given the retrieval task. Table 1 gives a high-level overview of the relevant corpora that we used. Each of them will be described in more detail in the three following sections. Depending on the strength of tail dependency in the data, we will see varying improvements for the three inspected settings. Comparable as the scenarios appear, there seem to be significant underlying differences in the distribution of relevant documents that influence the benefit from the use of copulas. In Section 6, we will dedicate some room to a detailed investigation of when the use of copula-based retrieval models is most promising.

Table 1: Overview of experimental corpora.

ID	# docs	# topics	# labels	year
Blogs08	1.3M	100	38.2k	2008
Delicious	339k	180	3.8k	2012
ODP	22k	30	1k	2009

4.1 Opinionated blogs

When conducting marketing analyses for businesses, researching customer reviews of products or gauging political trends based on voter opinions, it can be desirable to focus the search process on subjective, non-factual documents. The Text REtrieval Conference (TREC) accounted for this task within the confines of their Blog Track between the years 2006 and 2010 [32]. The aim of the task is to retrieve blog feeds that are both topically relevant and opinionated. Our experimental corpus for this task is the Blogs08 collection specifically created for the venue. The dataset consists of 1.3 million blog feeds and is annotated by more than 38k manually created labels contributed by NIST assessors.

Each document is represented as a two-component vector $U_{rel}^{(2)}$. The first component refers to the document's topical relevance given the query and the second represents its degree of opinionatedness. In order for a document to be considered relevant according to the judges' assessments, it has to satisfy both conditions. Topical relevance was estimated by a standard BM25 model and opinionatedness was determined using the output of a state-of-the-art open source classifier [1]. After an initial evaluation of the domain, we chose Clayton copulas (Equation 2) to represent the joint distribution of topicality and opinionatedness. Table 2 shows a juxtaposition of performance scores for the baselines as well as the various copula methods. The highest observed performance per metric is highlighted by the use of bold typeface, statistically significant improvements (measured by means of a Wilcoxon signed-rank test at $\alpha = 0.05$-level) over all competing approaches are denoted by an asterisk. Of the baseline methods, the score product PROD performs best. However, introducing the use of copulas, we observe that the highest performance was achieved using the *CPOS* copula,

Table 2: Copula-based relevance estimation performance for opinionated blogs ($k = 2$).

Method	P@5	P@10	p@100	BPREF	MRR	MAP
PROD	**0.413**	0.360	0.181	0.289	**0.692**	0.275
SUM	0.400	0.333	0.154	0.255	0.689	0.238
LIN	0.387	0.333	0.162	0.262	0.689	0.245
CPOS	**0.413**	**0.400***	**0.182**	**0.306***	**0.692**	**0.287***
CNEG	0.373	0.373	0.181	0.290	0.545	0.245
CODDS	0.373	0.360	**0.182**	0.283	0.544	0.242

Table 3: Copula-based relevance estimation performance for personalized bookmarks ($k = 2$).

Method	P@5	P@10	p@100	BPREF	MRR	MAP
PROD	0.084	0.079	**0.011**	0.051	0.192	0.043
SUM	0.095	0.095	**0.011**	0.071	0.192	0.055
LIN	0.126	**0.100***	**0.011**	0.077	**0.219***	0.063
CPOS	0.105	0.068	0.01	0.056	0.190	0.047
CNEG	**0.137***	0.090	0.010	**0.079***	0.184	**0.065**
CODDS	0.116	0.074	0.01	0.066	0.202	0.058

which gave statistically significant gains in MAP, Bpref and precision at rank 10 over all the baseline methods.

At this point, we revisit the example query (Topic 1171) that was discussed in the introduction and depicted in Figure 1. For this topic, we observed a clear non-linear dependency structure alongside a lower-than-average linear combination performance of $AP = 0.14$. When applying CPOS to the topic, however, we obtain $AP = 0.22$, an improvement of over 50%.

4.2 Personalized bookmarks

Finding and re-finding resources on the Internet are frequently accompanied and aided by bookmarking. What started as a local in-browser navigation aid, has in recent years become an active pillar of the social web society. Collaborative bookmarking platforms such as *Delicious*, *Furl*, or *Simpy* allow users to maintain an online profile along with bookmarks that can be shared among friends and collaboratively annotated by the user community. Research into tagging behaviour [2] found that a significant amount of the tags assigned to shared media items and bookmarks are of subjective nature and do not necessarily serve as objective topical descriptors of the content. This finding suggests that bookmarking has a strong personal aspect which we will cater for in our experiment. Vallet *et al.* [50] compiled a collection of more than 300k Delicious bookmarks and several million tags to describe them. For a share of 3.8k bookmarks and 180 topics, the authors collected manual relevance assessments along two dimensions, topical relevance of the bookmark given the topic and personal relevance of the bookmark for the user. This dataset is one of the very few corpora whose personalized relevance judgements were made by the actual users being profiled. We conduct a retrieval experiment in which we estimate topical and personal relevance for each document and use Gumbel copula models to model the joint distribution of facets. The set of relevant documents comprises only those bookmarks that satisfy both criteria and were judged relevant in terms of topicality and personal relevance. Table 3 shows an overview of the resulting retrieval performances. *CNEG* stands out as the strongest copula-based model but the overall ranking of systems depends on the concrete metrics evaluated. For some metrics such as precision at rank 10 and MRR, the linear combination baseline prevails, BPREF and precision at 5 documents favour CNEG.

4.3 Child-friendly websites

The third application domain that we will inspect is concerned with the retrieval of child-friendly websites. Children, especially at a young age, are an audience with specific needs that deviate significantly from those of standard web users. Even for adult users it has been shown that focussing the retrieval process on material of appropriate reading level can benefit user satisfaction [12]. In the case of children, this tendency can be expected to be even more pronounced since young users show very different modes of interaction with search engines that reflect their specific cognitive and motor capabilities [16]. Consequently, dedicated web search engines for children should focus their result sets on topically relevant, yet age-appropriate documents. [17] constructed a corpus of 22k web pages, 1,000 of which were manually annotated in terms of topical relevance towards a query as well as the document's likelihood of suitability for children. According to the authors, the class of suitable documents encompasses those pages that were topically relevant for children, presented in a fun and engaging way and textually not too complex to be understood. In our retrieval experiment, we account for both criteria and require documents to be both on topic as well as suitable for children in order to be considered relevant. Table 4 gives an overview of the resulting retrieval performance. In this setting, the various copula models show comparable result quality as the non parametric baselines. Linear combinations with empirically learned weights, however, were consistently the strongest method. We intend to explore the reasons for this in future work. However we note that the distribution of child-suitable ratings has a very large mode at zero, with only a small number of non-zero scores taking a limited number of possible discrete values - limiting the amount of useful dependency information available that copulas could exploit.

Table 4: Copula-based relevance estimation performance for child-friendly websites ($k = 2$).

Method	P@5	P@10	p@100	BPREF	MRR	MAP
PROD	0.240	0.143	0.051	0.221	0.349	0.196
SUM	0.246	0.157	0.052	0.213	0.340	0.200
LIN	**0.320***	**0.187***	**0.071***	**0.275***	**0.357**	**0.235***
CPOS	0.238	0.140	0.053	0.215	0.351	0.200
CNEG	0.242	0.140	0.048	0.223	0.349	0.194
CODDS	0.241	0.143	0.052	0.220	0.349	0.196

5. SCORE FUSION

Previously, we investigated the usefulness of copulas for modelling multivariate document relevance scores based on a number of (largely) orthogonal document quality criteria. Now, we will address a different, closely related problem: *score fusion* (also known as an instance of data fusion). In this setting, rather than estimating document quality from the documents, we attempt to combine the output of several independent retrieval systems into one holistic ranking. This challenge is often encountered in the domains of metasearch or search engine fusion. To evaluate the score fusion performance of copula-based methods, we use historic submissions

to the TREC Adhoc and Web tracks. We investigate 6 years of TREC (1995 - 2000) and fuse the document relevance scores produced by several of the original participating systems. Intuitively, this task closely resembles the previously addressed relevance estimation based on individual document properties. In practice, as we will show, the scenario differs from direct relevance estimation in that retrieval systems rely on overlapping notions of document quality (e.g., a variant of *tf/idf* scoring) and are therefore assumed to show stronger inter-criteria dependencies than individual facets of document quality might. Systematically, however, we address a set of document-level scores $U_{rel}^{(k)}$, originating from k retrieval systems, exactly in the same way as we did document quality criteria in the previous section.

As performance baselines, we will rely on two popular score fusion schemes, *CombSUM* and *CombMNZ*[19]. CombSUM adds up the scores of all k constituent retrieval models and uses the resulting sum as a new document score. CombMNZ tries to account for score outliers by multiplying the cross-system sum by $NZ(U)$, the number of non-zero constituent scores.

$$CombSUM(U_{rel}) = \sum_{i=1}^{k} u_{rel,i}$$

$$CombMNZ(U_{rel}) = NZ(U_{rel}) \sum_{i=1}^{k} u_{rel,i}$$

We introduce statistically principled, copula-based extensions of these established baseline methods: corresponding to CombSUM and CombMNZ, we define *CopSUM* and *CopMNZ* that normalize the respective baseline methods by the non-relevance copula.

$$CopSUM(U_{rel}, U_{non}) = \frac{\sum_{i=1}^{k} u_{rel,i}}{C_{non}(U_{non})}$$

$$CopMNZ(U_{rel}, U_{non}) = \frac{NZ(U_{rel}) \sum_{i=1}^{k} u_{rel,i}}{C_{non}(U_{non})}$$

Due to the close relationship to the baseline methods, the effect of introducing copulas is easily measurable. Based on empirical evidence, we employ Clayton copulas to estimate $C_{non}(U_{non})$.

Table 5 compares the baselines and copula methods in terms of MAP gain over the best, worst and median historic system run that were fused. Each performance score is averaged over 200 repetitions of randomly selecting k individual runs with k ranging from 2 to 10 for each year of TREC. Statistically significant improvements over the respective baseline method, *i.e.* of CopSUM over CombSUM and CopMNZ over CombMNZ, are determined by a Wilcoxon signed-rank test at $\alpha = 0.05$ level and are denoted by an asterisk.

Regarding the baseline methods, CombSUM and CombMNZ perform equally well on average, but with a clear dataset bias. On TREC 4, 8 and 9, CombSUM performs consistently better than CombMNZ. For TREC 5, 6 and 7, the inverse is true. With the exception of TREC 4, the fused rankings do not match the performance of the single strongest run that contributed to the fusion.

Introducing the copula methods led to consistent improvements over their non-copula baseline counterparts. In 104 out of 168 cases, the copula-based fusion methods gave statistically significant gains, with only 14 out 168 performing worse than the corresponding baseline method. The copula-based methods achieved, on average, 7% gains over the corresponding baseline when comparing to the strongest fused system, 4% gain on median systems and 2% gain on the weakest systems.

Fusion robustness

There are significant differences in fusion effectiveness between individual editions of TREC. Comparing TREC 4 and TREC 6, for example, we observe that TREC 6 fusion results typically showcase performance losses in comparison to the best original run and very high gains for the weakest systems. We seek an explanation in the imbalance in performance of the original systems. Very weak systems have the potential of decreasing the overall quality of the fused result list by boosting the scores of non-relevant documents. As the number of very weak systems increases, so does the chance for performance losses introduced by fusion. When inspecting the number weak submissions (defined as having an MAP score that is at least 2 standard deviations lower than the average score across all participants) included in our fusion experiments, we find that, indeed, our TREC 6 sample includes ~27% more weak systems than that of TREC 4.

In order to further investigate the influence of weak runs on overall fusion performance and to measure the proposed methods' robustness against this effect, we turn to the 10-system fusion scenario and inject more and more weak systems among the regular ones. Figure 3 shows how the fusion improvement over the single strongest system of TREC 4 is affected as the number of weak submissions ranges from 0 to 9 out of 10. As before, each data point is an average across 200 fusions of randomly drawn runs. In the ideal setting, in which there are no weak systems, we note higher performance gains than in the uncontrolled scenario that was shown in Table 5. As the number of weak systems injected into the fusion increases, performance scores quickly drop. As noted earlier, CombSUM performs slightly better on TREC 4 than CombMNZ. This difference, however, is not further influenced by the number of weak systems. The copula-based fusion methods are more resistant to the influence of weak systems. We note the divide between copula-methods and baseline approaches growing as the number of weak systems increases. Each baseline system score is well-separated from the respective copula-based variant. Error bars in Figure 3 were omitted to prevent clutter.

6. DISCUSSION

In Section 4, we investigated three different domains in which we apply copulas to model the joint distribution of multivariate relevance scores. For each of these settings, we could observe varying degrees of usefulness of the proposed copula scheme. While for child-friendly web search, the linear baseline performed best, we achieved significant improvements in the opinionated blog retrieval setting. At this point, we investigate the reason for this seeming imbalance in performance gains in order to find a way of deciding for which problem domains the application of copulas is most promising.

One of the key properties of copulas is their ability to account for tail dependencies. Formally, tail dependence describes the likelihood that component $u_{rel,i}$ within the observation vector $U_{rel}^{(k)}$ will take on extremely high or low values,

Table 5: Score fusion performance based on historic TREC submissions. Evaluated in percentages of MAP improvements over the best, median, and worst original systems that were fused.

TREC 4	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-9.8	-	118	-4.2	20	1128	0.0	33.5	1709	3.0	39.6	2344	3.9	48.5	3116
CopSUM	-9.6*	-	116	-4.2	20.5*	1136	0.0	33.8*	1721	3.2*	40.0*	2350	4.0	49.2*	3125*
CombMNZ	-9.5	-	116	-5.4	18.3	1071	-1.1	31.6	1675	2.1	38.3	2310	3.6	48.0	3106
CopMNZ	-9.5	-	115	-5.5	18.2	1080	-1.0	31.9*	1689*	1.8	38.6*	2318*	3.8*	48.0	3117*

TREC 5	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-5.6	-	268	-10.6	12.5	614	-6.9	26.5	955	-5.3	34.3	1031	-5.6	40.1	1479
CopSUM	-5.2*	-	274*	-9.9*	13.0*	613	-6.7*	28.0*	972*	-4.9*	35.0*	1050*	-5.2*	43.3*	1503*
CombMNZ	-4.6	-	269	-6.7	17.4	652	-3.5	30.9	986	-2.5	38.2	1074	-3.3	43.5	1526
CopMNZ	-4.5	-	274*	-6.5	17.8*	667*	-3.1*	32.2*	991	-2.4	38.7*	1092	-3.0*	46.0*	1554*

TREC 6	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-18.5	-	486	-24.6	7.8	2235	-24.0	29.6	3950	-22.8	44.9	5585	-22.1	56.9	7685
CopSUM	-17.7*	-	471	-23.1*	9.1*	2279*	-22.9*	32.1*	4075*	-21.2*	48.3*	5699*	-20.8*	58.2*	7702
CombMNZ	-17.0	-	491	-18.6	15.5	2537	-18.1	38.8	4386	-16.7	55.0	6111	-17.3	65.0	8117
CopMNZ	-16.3*	-	490	-17.2*	17.4*	2601*	-17.9	40.5*	4458*	-16.7	59.6*	6202*	-16.4*	66.8*	8170

TREC 7	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-9.3	-	132	-16.2	6.2	303	-11.7	25.9	504	-12.8	30.0	708	-14.5	36.3	863
CopSUM	-9.4	-	145*	-15.8*	6.5*	321*	-11.1*	27.2*	538*	-12.3*	34.1*	734*	-13.8*	39.1*	877
CombMNZ	-8.8	-	130	-13.7	9.4	347	-10.1	28.1	538	-10.9	32.8	745	-13.1	38.5	891
CopMNZ	-8.8	-	139*	-13.3*	10.1*	363*	-10.2	30.5*	565*	-10.7	34.7*	786*	-12.4*	40.4*	922

TREC 8	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-15.9	-	475	-11.6	8.1	1188	-11.5	16.9	3194	-7.7	21.8	2739	-5.4	21.8	3372
CopSUM	-16.1	-	488	-10.1*	8.3	1201	-10.9*	16.7	3195	-7.3*	22.3*	2755	-4.3*	22.4	3397
CombMNZ	-17.2	-	421	-11.8	7.6	1273	-12.9	15.1	3209	-9.8	18.6	2660	-7.2	19.2	3266
CopMNZ	-17.3	-	447*	-11.2	7.9*	1292	-12.8	14.9	3216	-9.2*	19.7*	2685	-6.7*	20.5*	3301*

TREC 9	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-9.0	-	173	-14.9	20.4	473	-15.6	17.4	178	-21.3	18.9	202	-27.9	12.6	204
CopSUM	-8.5*	-	188*	-13.7*	21.2*	499*	-15.3	17.9	182	-20.9	19.2*	207	-26.6*	13.1*	206
CombMNZ	-11.0	-	155	-19.0	14.5	435	-17.4	14.4	172	-25.3	12.6	186	-32.7	4.7	184
CopMNZ	-10.7	-	167	-17.9*	16.0*	432	-17.1	14.7	176	-24.8*	13.0*	190	-30.4*	5.1*	187

as another component $u_{rel,j}$ with $i \neq j$ also takes an extreme value. The strength of this correlation in extreme regions is expressed by the *tail dependency indices* I_U and I_L for upper and lower tail dependency, respectively. Higher values of I signal stronger dependencies in the respective tail regions of the distribution.

$$I_U = P\{X_1 > F_i^{-1}(u_i) | X_2 > F_j^{-1}(u_j)\}$$

$$I_L = P\{X_1 \leq F_i^{-1}(u_i) | X_2 \leq F_j^{-1}(u_j)\}$$

The literature has brought forward a number of estimators of the tail indices. We use the R implementation of Frees *et al.*'s method [20].

Tail index estimates serve as good tools for separating domains where we are likely to observe performance gains (blog and bookmark retrieval) and those that do not match linear combination performance (child-friendly search). Based on the respective copula models that we fit to our observations, the blog retrieval ($I_L = 0.07$) and personalized bookmarking ($I_U = 0.49$) show moderate tail dependencies while the child-friendly web search task has no recognizable dependency among extrema ($I_L = I_U = 0$). Since the comparison of absolute tail index scores across observations is not meaningful, we are interested in a method to further narrow down the expected performance. To this end, we took a closer look at the actual data distribution, and investigated goodness-of-fit tests that are used to determine how well an assumed theoretical distribution fits the empirical observations. The higher the likelihood of our observations to have been generated by the copula models that we estimated, the higher resulting performance we can expect. We apply a standard Anderson-Darling test [3] to determine how well the observations are represented by the copula models. In the personalized bookmarking setting, we obtain $p = 0.47$ and for the blog data $p = 0.67$ for the null hypothesis of the observations originating from the present copula model. As we suspected based on the tail dependency strength, the child-friendly web search data only achieved a probability of fit of $p = 0.046$.

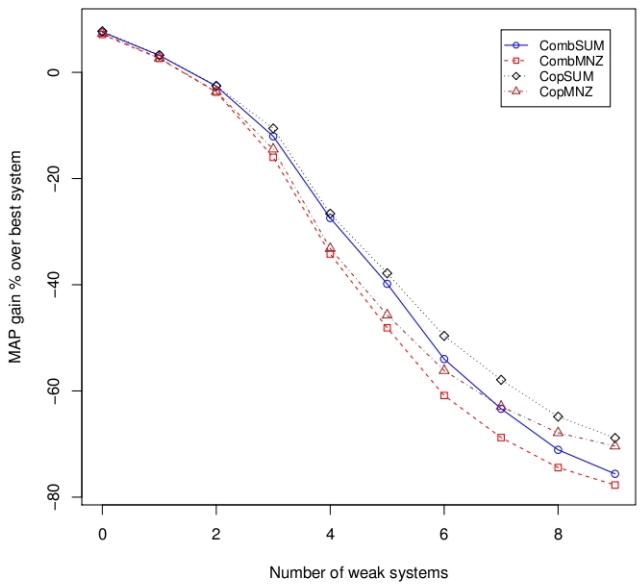

Figure 3: Performance in terms of MAP when $0 \ldots 9$ out of 10 fused original systems are weak.

To summarize, in this section, we have shown how a combination of tail dependence indices and goodness-of-fit tests can be used to help differentiate between domains that may benefit from copula-based retrieval models and those that may not.

7. CONCLUSION

In this work we introduced the use of *copulas*, a powerful statistical framework for modeling complex dependencies, for information retrieval tasks. We demonstrated the effectiveness of copula-based approaches in improving performance on several standard IR challenges. First, we applied copulas to the task of multivariate document relevance estimation, where each document is described by several potentially correlated relevance criteria. We learned and evaluated copula models for three different IR tasks, using large-scale standard corpora: (1) opinionated blog retrieval; (2) personalized social bookmarking; and (3) child-friendly web search, obtaining significant improvements on the first two of these tasks. Second, we introduced copula-based versions of two existing score fusion methods, COMB-Sum and COMB-MNZ, and showed that these improve the performance of score fusion on historic TREC submissions, in terms of both effectiveness and robustness, compared to their non-copula counterparts. Finally, we investigated the performance differences of copula models between different domains, and proposed the use of tail dependency indices and goodness-of-fit tests to understand the likely effect of using copulas for a given scenario.

In future work, there are a number of interesting challenges remaining in applying copula-based models to information retrieval. **(1)** The independence assumption between individual terms in queries and documents is a long-standing simplification in document and language modelling. Most attempts at incorporating more powerful dependency models into the retrieval process resulted in limited performance improvements at best. We would like to investigate the use of copulas in order to more realistically approximate the complex underlying term dependency structure. **(2)** During our investigation of the blog retrieval scenario, we encountered examples of non-linear multivariate distributions of relevance and briefly pointed out the different correlation regimes that exist within the joint distribution. While the current single-copula models have been shown to outperform linear combination models at capturing such structures, we would like to proceed to inspecting mixture models in which individual copulas account for certain data ranges to represent the underlying regimes better than a single holistic model could. **(3)** This work represents an exploratory study that aims to introduce the copula framework to the information retrieval community. For reasons of simplicity and brevity, it is based on data-driven estimation of copula parameters θ. It would, however, be interesting to build on the large body of previous work on formal modelling of the probability of relevance, to derive custom information retrieval copulas from the assumed distribution of relevance among documents.

8. REFERENCES

[1] Alias-i. LingPipe 3.9.2. http://alias-i.com/lingpipe, 2013.

[2] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *SIGCHI 2007*. ACM.

[3] TW Anderson and D.A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49, 1954.

[4] Avi Arampatzis and Stephen Robertson. Modeling score distributions in information retrieval. *Information Retrieval*, 2011.

[5] J.A. Aslam and M. Montague. Bayes optimal metasearch: a probabilistic model for combining the results of multiple retrieval systems (poster session). In *Proceedings of SIGIR 2000*, pages 379–381. ACM.

[6] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of SIGIR 2006*, pages 43–50. ACM.

[7] G. Bordogna and G. Pasi. A model for a SOft Fusion of Information Accesses on the web. *Fuzzy Sets and Systems*, 148(1):105–118, 2004.

[8] P. Borlund. The concept of relevance in IR. *JASIST*, 2003.

[9] J.P. Bouchaud and M. Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge University Press, 2003.

[10] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96. ACM, 2005.

[11] A. Charpentier, J.D. Fermanian, and O. Scaillet. The estimation of copulas: Theory and practice. *Copulas: From theory to Application in Finance. Risk Publications*, 2007.

[12] K. Collins-Thompson, P.N. Bennett, R.W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *CIKM 2011*. ACM.

[13] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *Proceedings of SIGIR 2005*, pages 416–423. ACM.

[14] Ronan Cummins. Measuring the ability of score distributions to model relevance. In *Information Retrieval Technology*. Springer, 2011.

[15] C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional relevance: A new aggregation criterion. *ECIR 2009*.

[16] A. Druin, E. Foss, L. Hatley, E. Golub, M.L. Guha, J. Fails, and H. Hutchinson. How children search the internet with keyword interfaces. In *Proceedings of the 8th International Conference on Interaction Design and Children*, pages 89–96. ACM, 2009.

[17] C. Eickhoff, P. Serdyukov, and A.P. de Vries. A combined topical/non-topical approach to identifying web sites for children. In *WSDM 2011*. ACM.

[18] P. Embrechts, F. Lindskog, and A. McNeil. Modelling dependence with copulas and applications to risk management. *Handbook of heavy tailed distributions in finance*, 8(329-384):1, 2003.

[19] E. Fox and J. Shaw. Combination of multiple searches. *NIST Special Pub.*, 1994.

[20] E.W. Frees and E.A. Valdez. Understanding relationships using copulas. *North American actuarial journal*, 2(1), 1998.

[21] S. Gerani, C.X. Zhai, and F. Crestani. Score transformation in linear combination for multi-criteria relevance ranking. *ECIR 2012*.

[22] S.P. Harter. Psychological relevance and information science. *JASIS*, 43(9):602–615, 1992.

[23] W. Höffding. Scale-invariant correlation theory. *Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universität Berlin*, 5(3):181–233, 1940.

[24] X. Huang and W.B. Croft. A unified relevance model for opinion retrieval. In *Proceeding of CIKM 2009*, pages 947–956. ACM.

[25] Evangelos Kanoulas, Keshi Dai, Virgil Pavlu, and Javed A Aslam. Score distribution models: assumptions, intuition, and robustness to score manipulation. In *SIGIR 2010*. ACM.

[26] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[27] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR*. ACM, 2002.

[28] V. Lavrenko and W.B. Croft. Relevance based language models. In *Proceedings of SIGIR 2001*, pages 120–127. ACM.

[29] V. Lavrenko and W.B. Croft. Relevance models in information retrieval. *Language modeling for information retrieval*, pages 11–56, 2003.

[30] T.Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 2009.

[31] W. Lu, S. Robertson, and A. MacFarlane. Field-weighted xml retrieval based on bm25. *Advances in XML Information Retrieval and Evaluation*, pages 161–171, 2006.

[32] C. Macdonald, R.L.T. Santos, I. Ounis, and I. Soboroff. Blog track research at trec. In *SIGIR Forum 2010*. ACM.

[33] R. Manmatha, Toni M. Rath, and Fangfang Feng. Modeling score distributions for combining the outputs of search engines. In *SIGIR 2001*.

[34] S. Mizzaro. Relevance: The whole history. *JASIS*, 1997.

[35] M. Montague and J.A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of CIKM 2002*, pages 538–548. ACM.

[36] M. Montague and J.A. Aslam. Relevance score normalization for metasearch. In *CIKM 2001*. ACM.

[37] A. Onken, S. Grünewälder, M.H.J. Munk, and K. Obermayer. Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flashlight transformation. *PLoS computational biology*, 5(11):e1000577, 2009.

[38] M. Persin, J. Zobel, and R. Sacks-Davis. Filtered document retrieval with frequency-sorted indexes. *JASIS*, 47(10):749–764, 1996.

[39] J.M. Ponte and W.B. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR 1998*, pages 275–281. ACM.

[40] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *SIGKDD*, pages 239–248. ACM, 2005.

[41] B. Renard and M. Lang. Use of a gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Advances in Water Resources*, 30(4):897–912, 2007.

[42] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *CIKM 2004*.

[43] S.E. Robertson, S. Walker, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. Gaithersburgh, MD, 1994.

[44] T. Saracevic. Relevance reconsidered. In *Conference on Conceptions of Library and Information Science*, 1996.

[45] L. Schamber, M.B. Eisenberg, and M.S. Nilan. A re-examination of relevance: toward a dynamic, situational definition. *IPM*, 26(6):755–776, 1990.

[46] T. Schmidt. Coping with copulas. *Risk Books: Copulas from Theory to Applications in Finance*, 2007.

[47] C. Schoelzel, P. Friederichs, et al. Multivariate non-normally distributed random variables in climate research–introduction to the copula approach. *Nonlin. Processes Geophys.*, 15(5):761–772, 2008.

[48] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8(1):11, 1959.

[49] T. Tsikrika and M. Lalmas. Combining evidence for relevance criteria: a framework and experiments in web retrieval. *ECIR 2007*.

[50] D. Vallet and P. Castells. Personalized diversification of search results. In *SIGIR 2012*. ACM.

[51] C.C. Vogt and G.W. Cottrell. Fusion via a linear combination of scores. *Information Retrieval*, 1(3):151–173, 1999.

[52] S. Wu and F. Crestani. Data fusion with estimated weights. In *CIKM 2002*. ACM.