# Robust Statistical Methods in Web Retrieval

Carsten Eickhoff[1] and Arjen P. de Vries[2]
[1]Department of Computer Science, ETH Zurich, Switzerland
[2]Information Foraging Lab, Radboud University Nijmegen, The Netherlands

Information retrieval systems rely on multitudes of individual features in order to determine the ranking of documents for a given user and query combination. 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 scores. To address these issues, we employ copulas, a family of robust statistical methods, introducing their formal background and empirically demonstrating their merit in a number of settings, including ranking, score fusion and language modelling.

DOI: 10.1145/2857659.2857663 http://doi.acm.org/10.1145/2857659.2857663

## 1. INTRODUCTION

Information retrieval systems enable searching in and browsing through massive collections of documents. Considering the scale and growth rate of the Internet, search engines have become indispensable in modern days. In response to user queries, they 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 [Robertson et al. 2004]. 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 [Collins-Thompson et al. 2011] or suitability for the user's age group [Eickhoff et al. 2011] have begun influencing the ranking process.

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 may appear independent for most 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 [Macdonald et al. 2010], 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 dimensions. We can note an apparent correlation in the lower left part of the graph that weakens as scores

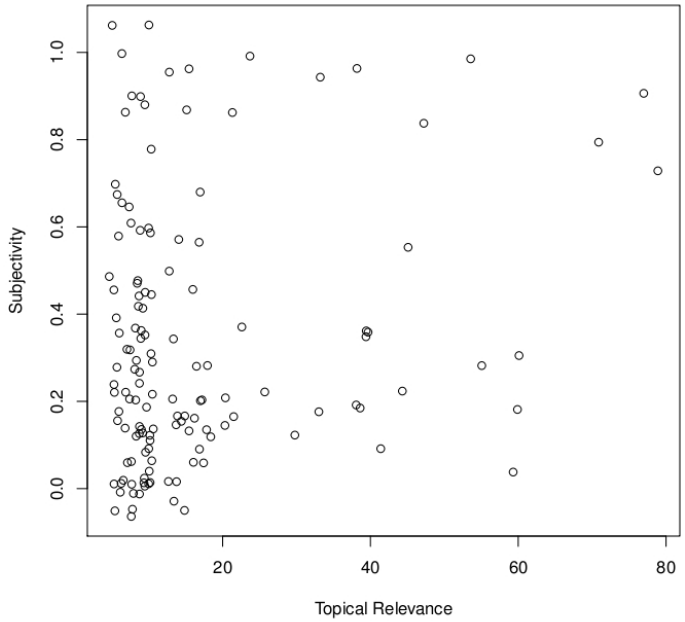


Fig. 1: Distribution of bivariate relevance scores for TREC 2010 Blog Track Topic 1171, "mysql".

increase. To underline this, we computed Pearson's $\rho$ 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 $\lambda$, Topic 1171 receives an average precision of only 0.14, well below the method's average across all topics of 0.25.

While the machine learning, information retrieval, data mining and natural language processing communities have significant expertise in estimating relevance criteria such as the document's topical relevance in isolation, the commonly applied combination schemes have tended to be *ad hoc*; ignoring the problem of modeling complex, multi-dimensional dependencies. In practice, they follow statically weighted linear combinations with empirically determined mixture parameters [Robertson et al. 2004] or deploy sophisticated learning to rank techniques [Liu 2009] that offer only limited insight to humans. Ideally, we would demand realistic, yet formally-grounded combination schemes that can lead to results that are both effective and supported by a human-interpretable justification.

The field of quantitative risk management faces a similar challenge to combine distinct indicators in a human-interpretable way. Here, researchers have made extensive use of *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 ranges of the interacting distributions. On the stock market, they are used to describe the relationship between commodities that 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 [Bouchaud and Potters 2003]. Our experiments find a similar behavior of quasi-independent facets of document relevance; as discussed above, these may also influence each other in the case of extreme conditions.

This article aims to summarize a number of recent advances into using copulas for a range of information retrieval tasks by reviewing the relevant theoretical foundations and demonstrating their application in a number of standard scenarios such as the previously mentioned blog retrieval example.

## 2. FORMAL BACKGROUND

Our use of copulas follows a similar pattern, irrespective of the concrete application scenario. We cast our observations (i.e., individual features) as random vectors. The respective likelihoods of observing this vector given the copula describes our probability of relevance, used as the final ranking criterion. Before we move on to discussing a number of applications, let us briefly introduce some necessary notation and formal background. For a more comprehensive overview of the copula framework, please refer to [Embrechts et al. 2003].

Let $X$ be a $k$-dimensional random vector of observations that we wish to use as input to our copula model:

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

The copula allows us to model the likelihood of observing $X$ by offering computationally efficient approximations to the true joint probability distribution in the high-dimensional space of cardinality $k$. As a first step, copulas require input scores $U^k$ to be uniformly distributed in the $[0, 1]$ interval. We can achieve this by defining a set of transformations $F(X)$ between the raw marginal observations $X$ and their normalized equivalents on the unit cube $U$.

$$U^k = (u_1, u_2, \ldots, u_k) = F(X) = (f_1(u_1), f_2(u_2), \ldots, f_k(u_k))$$

For each of our $k$ dimensions, a readily available example of such a function is given by the empirical distribution function $\hat{f}$ that asymptotically approximates the true underlying distribution function $f$ as the number of samples $n$ increases:

$\hat{f}(t) = \frac{1}{n} \sum 1 \{x_i \leq t\}$

The cumulative distribution function $C$ for all explicitly given copulas is fully defined in terms of a generator $\psi$ and its inverse $\psi^{-1}$:

$$C(u_1, u_2, \ldots, u_k) = \psi^{-1}(\psi(u_1) + \psi(u_2) + \ldots + \psi(u_k))$$

Many concrete instantiations of such copula functions have been proposed. Each so-called copula *family* defines their own generator and inverse. Previous investigations found Gumbel copulas to be an adequate choice in Web retrieval settings. Their generators are given in the following form:

$$\psi^{-1}(t) = exp(-t^{\frac{1}{\theta}})$$
$$\psi(t) = (-log(t))^{\theta}$$

The resulting distribution function for a $k = 2$-dimensional Gumbel copula is:

$$C(u_1, u_2) = exp(-((-log(u_1))^{\theta} + (-log(u_2))^{\theta})^{\frac{1}{\theta}})$$

As we can see, there is a single parameter $\theta$ that allows us to control the strength of dependency between the individual marginal observations $u$. If we, for example, set $\theta = 1$, our distribution function defaults to the case of conditional independence:

$$C_{\theta=1} = exp(-(-log(u_1)) + (-log(u_2))) = u_1 * u_2$$

Any choice of $\theta > 1$ results in an increasing degree of conditional dependency between the $k$ dimensions of our observation. While this serves for elegant models controlled by only a single degree of freedom, there are some scenarios, especially for large $k$, in which this approach may not be optimal since it assumes identical dependency structures between all dimensions. Instead of combining all dimensions in a single step, we can alternatively define a nested hierarchy of multiple copulas that estimate joint distributions for sub sets of the full feature space and subsequently combine scores until one global model is obtained [Eickhoff and de Vries 2014]. Generally, an example of a fully nested copula with $k$ dimensions is given by:

$$C_0(u_1, C_1(u_2, C_2(\ldots, C_{k-2}(u_{k-1}, u_k))))$$

By means of the structure of the nesting "*tree*", nested copulas can explicitly model which dimensions depend on each other directly. Instead of the global $\theta$ parameter discussed earlier, each of the constituent copulas defines their respective $\theta_i$, determining the strengths of these (per-dimension) dependencies. This mechanism gives nested copulas a theoretical

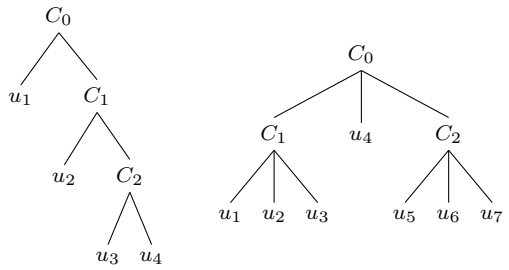


Fig. 2: Examples of fully nested (left) and partially nested (right) copulas.

advantage in flexibility over their non-nested counterparts. Figure 2 shows a fully nested copula with $k-1$ copula modelling steps (left) and a conceptual example of a partially nested copula (right).

As a final step, for ranking and classification applications, we require point estimates of probability density $c(U)$ which are obtained via partial integration:

$$c(U) = \frac{\partial^k}{\partial_1 ... \partial_k} C(U)$$

Now, with the essential steps of the copula modelling process in place, let us consider three concrete applications of copulas to standard retrieval tasks.

## 3. GENERATIVE RELEVANCE MODELLING

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 [Macdonald et al. 2010]. 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 $U^{(2)}$. 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 [Alias-i. 2015].

We begin by separately estimating the probability of relevance $P^{(2)}_{rel}(d)$ and non-relevance $P^{(2)}_{non}(d)$ for a document $d$, under each of the $k = 2$ criteria (dimensions) – in this case, topicality, and subjectivity. Next, we assume random observations $U$ to derive from either of these distributions and train two distinct copulas, $C_{rel}$ and $C_{non}$.
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)** $CPOS(U)$ multiplies the independent likelihood of observing $U$ under the relevance copula $C_{rel}$, capturing only dependencies between the likelihoods of relevance. **(2)** $CNEG(U)$ normalizes the probability of relevance by the non-relevance copula $C_{non}$, capturing only the dependencies between the likelihoods of non-relevance. **(3)** $CODDS(U)$, finally, multiplies the probability of relevance by the ratio of the two copulas, modelling simultaneously the dependencies between both previous notions.

$$CPOS(U) = c_{rel}(U) \prod_{i=1}^{k} u_i$$

$$CNEG(U) = \frac{\prod_{i=1}^{k} u_i}{c_{non}(U)}$$

$$CODDS(U) = \frac{c_{rel}(U)}{c_{non}(U)} \prod_{i=1}^{k} u_i$$

As performance baselines, we compare to three popular combination methods from the literature: **(1)** $SUM(U)$ sums up the relevance scores across all criteria $k$ and uses the sum as the final ranking criterion [Fox and Shaw 1994]. **(2)** $PROD(U)$ builds the product across all constituents. Probabilistically, this combination scheme assumes independence across all criteria and can be expected to be naïve in some settings where dependence is given. **(3)** Weighted linear combinations $LIN_\Lambda(U)$ build a weighted sum of constituents $u_i$ with mixture parameters $\lambda_i$ optimized by means of a parameter sweep with step size $0.1$ [Vogt and Cottrell 1999]. It should be noted that all optimizations and parameter estimations, both for the baselines as well as for the copula models are conducted on the training portion of the corpus.

$$SUM(U) = \sum_{i=1}^{k} u_i$$

$$PROD(U) = \prod_{i=1}^{k} u_i$$

$$LIN_\Lambda(U) = \sum_{i=1}^{k} \lambda_i u_i$$

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 our copula models, we observe that the highest performance was achieved using

Table I: 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 |

the $CPOS$ copula, 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 with its bivariate relevance scores 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%. For more case studies of generative relevance modelling based on copulas, please refer to [Eickhoff et al. 2013]

## 4. DATA 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 though closely related problem: *score fusion* (a specific instance of the more general problem 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, a challenge encountered in practice in the domain of federated search 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 the entries to TREC 4 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., close variations 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 $U^{(k)}$, 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 $CombMNZ$[Fox and Shaw 1994]. 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.

$$CombSUM(U) = \sum_{i=1}^{k} u_i$$

$$CombMNZ(U) = NZ(U) \sum_{i=1}^{k} u_i$$

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.

$$CopSUM(U) = \frac{\sum_{i=1}^{k} u_i}{c_{non}(U)}$$

$$CopMNZ(U) = \frac{NZ(U) \sum_{i=1}^{k} u_i}{c_{non}(U)}$$

Due to the close relationship to the baseline methods, the effect of introducing copulas is easily measurable.

Table II compares the baselines and copula methods in terms of MAP gain over the best, worst and median historic system run that was 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.

Table II: Score fusion performance based on TREC 4 submissions. In percentages of MAP improvements over the best, median, and worst original system that was 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* |

CombSUM performs consistently better than CombMNZ. As the number of fused systems increases, the relative quality of the fusion methods improves, leading to significant improvements even over the best individual submissions in several cases. For an overview of fusion results for TREC editions 5 to 9, as well as a dedicated analysis of fusion robustness, please refer to [Eickhoff et al. 2013], where for 104 out of 168 compared cases, the copula-based fusion methods gave statistically significant gains, with only 14 out 168 performing worse than the corresponding baseline method. Copula-based methods achieved, on average, a gain of 7% over the corresponding baseline when comparing to the strongest fused system, a gain of 4% over median systems, and, a gain of 2% over the weakest systems.

## 5. LANGUAGE MODELLING

A traditional unigram language model describes the likelihood of observing a string of text $S$ under a given topical class $t$ as the product of the individual likelihoods of each term:

$$P(S|t) = P(s_1|t)P(s_2|t)\dots P(s_{|S|}|t)$$

The same can be achieved under the copula framework by considering the class-conditional probabilities of observing individual terms $s_1, s_2, \ldots$ as our marginal observations, making the dimensionality of our copula $k = |S|$:

$$P(S|t) = c_{\theta=1}(P(s_1|t), P(s_2|t), \ldots P(s_{|S|}|t))$$

By choosing $\theta = 1$, we ensure conditional independence between the marginal term observation likelihoods, giving us the standard unigram language model. As we however increase $\theta$, the strength of dependency between the individual terms increases. This ability to account for term dependence makes the copula framework a powerful alternative to the standard language modelling scheme. At this point, any setting of $\theta$ globally describes the relationship between all terms. In practice, however, we much rather want a select few terms to depend on each other, while the majority of terms indeed occur independently.

This is easily achieved by using nested copulas, as described earlier. At this point, the final missing component in our language modelling scheme is a way to determine the concrete settings of $\theta$ for each pair of terms. To this end, we define conditional dependency in terms of frequency of co-occurrence in a document corpus and rely on the point-wise mutual information between terms $s_1$ and $s_2$. For an in-depth investigation of alternative co-occurrence metrics, please refer to [Eickhoff et al. 2015].

$$PMI(s_1, s_2) = log_2 \frac{P(s_1,s_2)}{P(s_1)P(s_2)}$$

Finally, the dependency parameter $\theta_{s_1,s_2}$ is defined on the basis of the collection-wide metric mean ($\mu_{PMI}$) across all potential term pairs. All those pairs of higher-than-average co-occurrence frequency are assigned values of $\theta$ proportionally to their relative co-occurrence rate. Since $\theta$ is defined in the range $[1, \infty)$ and the resulting scores scale in a non-linear fashion, there is no need to further address or remove outlier pairs of extremely high frequency.

$$\theta_{s_1,s_2} = \begin{cases} \frac{PMI(s_1,s_2)}{\mu_{PMI}} & if\ PMI(s_1, s_2) > \mu_{PMI} \\ 1 & else \end{cases}$$

To empirically test our copula-based language modelling scheme, we investigate its performance at the task of adhoc document retrieval. Instead of modelling the likelihood of observing a given document under a topic specific language model, we will now establish one distinct model per document and compare their respective likelihoods of having generated the query $q$.

$$P(rel|q, d) \approx P(q|d)$$

$$P(q|d)_{indep} = \prod_{i=1}^{|q|} P(s_i|d)$$

Table III: Performance on ClueWeb'12 and TREC 2013 adhoc topics at a threshold of 20 retrieved documents.

| Model | Precision | Recall | $F_1$ | MAP |
|---|---|---|---|---|
| Unigram LM | 0.31 | 0.22 | 0.26 | 0.41 |
| Bigram LM | 0.34 | 0.26 | 0.3 | 0.45 |
| SenTree | 0.35 | 0.28 | 0.31 | 0.47 |
| MRF | 0.38 | 0.31 | 0.34 | 0.51 |
| Copula LM | **0.41*** | **0.35*** | **0.38*** | **0.52** |

$$P(q|d)_{cop} = c_d(s_1, s_2, \ldots, s_n)|s \in q$$

For our experimental comparison, we rely on the widely used ClueWeb'12 corpus, a collection of 730 million authentic Web documents. Our 50 topics originate from TREC's 2013 Adhoc retrieval task [Collins-Thompson et al. 2014]. We contrast our method's performance with a number of established as well as state-of-the-art baselines such as standard unigram and bigram language models, Nallapati's sentence trees [Nallapati and Allan 2002], as well as the Markov Random Field model [Metzler and Croft 2005], and apply Laplace smoothing to all LM variants in order to account for previously unseen query terms. Table III details the respective performances obtained by the various methods in terms of precision, recall, $F_1$ and MAP, each computed at a cut-off rank of 20 retrieved documents. Statistically significant improvements over all baseline methods are indicated by the asterisk character. Statistical significance was tested using a Wilcoxon signed-rank test at $\alpha \leq 0.05$-level.

We can note that, due to their wider context, the classification performance of bigram language models significantly exceeds that of the lower-order model. The models that explicitly capture term dependence, Sentence trees and Markov Random Fields, show even higher classification performance. Our copula language model yields significant performance improvements across most metrics and baselines, although improvements over the MRF model were only significant for some of the considered metrics.

## 6. CONCLUSION

This article has summarized an ongoing line of work that applies copulas, a model family from the field of robust statistics, at a number of standard Web retrieval tasks. The model identifies a number of theoretical benefits: (1) Scale invariance via decoupling of marginals and dependency structure, (2) easy approximation of high-dimensional joint distributions (3) the ability to explicitly model non-linear (tail) dependencies. After a formal introduction, we demonstrated the method's adequacy for the tasks of generative relevance modelling as used in standard retrieval systems, fusion of multiple independent system scores, as well as n-gram language modelling, showing significant gains across all domains. A particularly compelling advantage lies in the convenient interpretability of copula models, making them a prime tool in settings such as academic research or industrial prototyping where frequent human inspection is inevitable.

## REFERENCES


ALIAS-I. 2015. LingPipe 3.9.2. http://alias-i.com/lingpipe.

BOUCHAUD, J.-P. AND POTTERS, M. 2003. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge University Press.

COLLINS-THOMPSON, K., BENNETT, P., DIAZ, F., CLARKE, C. L., AND VOORHEES, E. M. 2014. TREC 2013 Web Track Overview. *Michigan University Ann Arbor, Technical Report*.

COLLINS-THOMPSON, K., BENNETT, P. N., WHITE, R. W., DE LA CHICA, S., AND SONTAG, D. 2011. Personalizing Web Search Results by Reading Level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, 403–412.

EICKHOFF, C. AND DE VRIES, A. P. 2014. Modelling Complex Relevance Spaces with Copulas. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*. ACM, 1831–1834.

EICKHOFF, C., DE VRIES, A. P., AND COLLINS-THOMPSON, K. 2013. Copulas for Information Retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 663–672.

EICKHOFF, C., DE VRIES, A. P., AND HOFMANN, T. 2015. Modelling Term Dependence with Copulas. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

EICKHOFF, C., SERDYUKOV, P., AND DE VRIES, A. P. 2011. A Combined Topical/Non-topical Approach to Identifying Web Sites for Children. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*. ACM, 505–514.

EMBRECHTS, P., LINDSKOG, F., AND MCNEIL, A. 2003. Modelling Dependence with Copulas and Applications to Risk Management. *Handbook of Heavy Tailed Distributions in Finance 8,* 329-384, 1.

FOX, E. A. AND SHAW, J. A. 1994. Combination of Multiple Searches. *NIST Special Publication SP*, 243–243.

LIU, T.-Y. 2009. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval 3,* 3, 225–331.

MACDONALD, C., SANTOS, R. L., OUNIS, I., AND SOBOROFF, I. 2010. Blog Track Research at TREC. In *ACM SIGIR Forum*. Vol. 44. ACM, 58–75.

METZLER, D. AND CROFT, W. B. 2005. A Markov Random Field Model for Term Dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 472–479.

NALLAPATI, R. AND ALLAN, J. 2002. Capturing Term Dependencies using a Language Model based on Sentence Trees. In *Proceedings of the 11th International Conference on Information and Knowledge Management*. ACM, 383–390.

ROBERTSON, S., ZARAGOZA, H., AND TAYLOR, M. 2004. Simple BM25 Extension to Multiple Weighted Fields. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*. ACM, 42–49.

VOGT, C. C. AND COTTRELL, G. W. 1999. Fusion via a Linear Combination of Scores. *Information Retrieval 1,* 3, 151–173.


---


Carsten Eickhoff is a postdoctoral researcher at ETH Zurich. He obtained his Ph.D. from Delft University of Technology in the Netherlands. His research focuses on crowdsourcing, search personalization and formal models of multi-dimensional document relevance.

Arjen P. de Vries is professor of Information Retrieval at Radboud University Nijmegen in the Netherlands. He obtained his Ph.D. in computer science from the University of Twente. His main research focus concerns the question how search technologies should be combined with user interactions, and, more specifically, how users and systems may cooperate to take advantage of structured and unstructured information representations to improve the access to information.