Modelling Complex Relevance Spaces with Copulas

Carsten Eickhoff
Dept. of Computer Science
ETH Zurich, Switzerland
ecarsten@inf.ethz.ch

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

ABSTRACT

Modern relevance models consider a wide range of criteria in order to identify those documents that are expected to satisfy the user's information need. With growing dimensionality of the underlying relevance spaces the need for sophisticated score combination and estimation schemes arises. In this paper, we investigate the use of copulas, a model family from the domain of robust statistics, for the formal estimation of the probability of relevance in high-dimensional spaces. Our experiments are based on the *MSLR-WEB10K* and *WEB30K* datasets, two annotated, publicly available samples of hundreds of thousands of real Web search impressions, and suggest that copulas can significantly outperform linear combination models for high-dimensional problems. Our models achieved a performance on par with that of state-of-the-art machine learning approaches.

Categories and Subject Descriptors

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

Keywords

Relevance models; Multivariate relevance; Ranking; Probabilistic framework.

1. INTRODUCTION

To address users' information needs, modern retrieval systems return result lists ordered by decreasing values in estimated *relevance*. Considering only topicality, relevance has been successfully estimated by term overlap between queries and documents [9]. There is, however, a wide range of theoretical relevance frameworks according to which relevance goes beyond mere topicality and is a composite notion comprised of dimensions such as document recency, credibility or monetary cost. There are several examples of applications focusing on non-topical factors such as textual complexity [2] or suitability for children [4]. Depending on the context of the search session, such factors can have a strong influence on result lists. A topically highly relevant document that is not understandable due its complex syntactic structure or its excessive use of jargon should be considered of lower effective relevance. With respect to these developments, the traditional task shifts from a univariate ranking problem to a multivariate one with many, potentially independent, dimensions.

The state of the art in multidimensional relevance modelling is dominated by two popular approaches. On the one hand, linear score combinations deliver intuitively interpretable, yet simplistic results. On the other hand, sophisticated learning-to-rank models often show superior performance that can come at the cost of offering less direct insight to human investigators. The recently presented copula framework for information retrieval [3] tries to overcome these inherent limitations by presenting a model that is formally grounded in probability theory while at the same time enabling flexible fitting to complex real-world distributions of relevance. The model's ability to account for co-movements in extreme regions of the relevance scale, so-called *tail dependencies* makes it an especially powerful framework that exceeds the capabilities of strictly linear functions. The initial copula approach concentrated on rather simple, two-dimensional, models. In this paper, we expand the original work by investigating modifications to the framework, especially geared towards use in high dimensional settings.

The research presented in this paper is guided by two fundamental research questions: **(1)** How does the retrieval performance of copulas compare to that of established IR models in high-dimensional relevance spaces? **(2)** More specifically, can nested copulas provide better approximations of the true underlying distribution of relevance in high-dimensional settings?

This paper goes beyond the current state of IR literature by investigating the use of model families from the domain of robust statistics to information retrieval settings. Expanding on the recently proposed copula-based relevance models, we present a number of experiments in high-dimensional relevance spaces, motivating the use of formal, yet data-driven models even in complex settings that traditionally are reserved for machine-learned approaches.

2. RELATED WORK

"*Relevance*" is a central notion in IR theory and applications. Many theoretical relevance frameworks have been proposed, including for example, [10], or [8]. Despite different definitions of the concrete composition of relevance,

CIKM'14, November 3–7, 2014, Shanghai, China.

http://dx.doi.org/10.1145/2661829.2661925.

most of these frameworks agree on the fact that relevance is a more complex notion than could be expressed in the form of a single criterion. Applied retrieval model implementations, for a long time, tended to rely on weighted linear combinations of individual relevance scores, for example in the popular BM25F scheme [9]. Gerani *et al.* [6] applied non-linear transformations prior to the linear combination step. Their positive results motivate the need for models whose capabilities go beyond strictly linear dependency structures. Kraaij *et al.* [7] investigated the formal combination of independent relevance dimensions in the form of prior probabilities injected into n-gram language models. As an alternative to the previously presented formal approaches, industrial solutions often rely on machine learning techniques in order to infer optimal rankings based on a wide range of features [1].

In order to combine the strengths of theoretically grounded models and machine-learned rankers, we applied copulas [3], a model family from the domain of robust statistics for the task of relevance modelling. Traditionally, copulas have been used in domains such as the analysis of stock portfolio risk, or meteorology, in which multitudes of variables interact in potentially non-linear fashion. While the original paper exclusively investigates two-dimensional relevance spaces, this paper studies high-dimensional relevance settings with more than one hundred individual features. Additionally, we further refine the previous approach by using nested copulas, an expansion to the copula framework that aims especially at high-dimensional settings.

3. METHODOLOGY

In the following, we will begin with a necessarily brief overview of the copula framework and its key properties. For a more comprehensive overview of this powerful model family, please refer to adjunct resources such as the survey by Embrechts *et al.* [5].

3.1 Copulas

Each copula is a multivariate *cumulative distribution function* (*cdf*). Given

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

a k-dimensional random vector with continuous margins

$$F_i(x_i) = \mathbb{P}[X_i \leq x_i]$$

observations can be mapped to the unit cube $[0,1]^k$ as

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

A k-dimensional copula $\mathcal{C}$ describes the joint cumulative distribution function of the normalized random vector U.

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

There are three particularly interesting properties that motivate the use of copulas for settings such as relevance modelling: (1) Since observations u and dependency structures $\mathcal{C}(.)$ are separated, each component is more straightforward to estimate. (2) The explicit dependency model operates on the unit cube which makes the method inherently scale invariant. (3) The ability to represent tail dependencies allows for accurate models of non-linear interdependence between relevance dimensions.

Different copula families have individual properties, strengths and limitations. Please refer to Embrechts *et al.* [5] for more detail. In this paper, for reasons of brevity and scope, we focus on using Gumbel copulas. This choice is motivated theoretically by their ability to account for tail dependencies in both the upper and lower extremes of the scale, that other models lack. A dedicated set of experiments empirically supports our preference by showing Gumbel-family copulas to consistently and significantly outperform the competing methods in the candidate pool. The comparison was conducted on the training portion of the *MSLR-WEB10K* dataset and further considered Gaussian, Joe, Clayton and Frank copulas. Gumbel copulas are formally given by:

$$\mathcal{C}_{Gumbel}(U) = \exp(-(\sum_{i=1}^{k}(-\log(u_i))^{\theta})^{\frac{1}{\theta}})$$

The model involves a single degree of freedom, the parameter $\theta \in [1, \infty]$, expressing the strength of dependency across dimensions k. In order to model the probability of document relevance via copulas, we propose a modification of the method presented in [3]. Based on the training portion of our dataset, we build individual copulas C_r and C_n (with parameters θ_r and θ_n) for modelling relevant and non-relevant documents, respectively. The probabilities of relevance and non-relevance are estimated by the respective copula densities given observation vector U. The final ranking criterion is given by $OR(rel|U)$, the odds ratio of relevance according to the two copulas.

$$\begin{aligned} OR(rel|U) &= \frac{P(rel|U)}{P(non|U)} \\ &= \frac{\mathcal{C}_r(U)\prod u_i}{\mathcal{C}_n(U)\prod u_i} \\ &= \frac{\mathcal{C}_r(U)}{\mathcal{C}_n(U)} \end{aligned}$$

3.2 Nested Copulas

Previous work [3] modelled low-dimensional document relevance spaces with a single copula with k components, where k equalled the cardinality of the entire relevance space. While this is possible in high-dimensional spaces as well, alternative options offer additional degrees of freedom. The use of so-called *nested* copulas is one such method. Instead of combining all dimensions in a single step as described earlier, they allow for a nested hierarchy of copulas that estimate joint distributions for sub sets of the full relevance space and subsequently combine scores until one global model is obtained. Formally, fully nested copulas with k dimensions are given by

$$\mathcal{C}(U) = \mathcal{C}_0(u_1, \mathcal{C}_1(u_2, \mathcal{C}_2(\ldots, \mathcal{C}_{k-2}(u_{k-1}, u_k))))$$

By means of the structure of the nesting "*tree*", nested copulas can explicitly model which dimensions depend on each other directly. The respective θ_i parameters determine the strengths of these (per-dimension) dependencies. This mechanism gives nested copulas a theoretical advantage in flexibility over their non-nested counterparts. As an alternative approach to full nesting, partially nested copulas hierarchically combine subsets of dimensions. Figure 1 shows a fully nested copula with $k-1$ copula modelling steps (left) and a conceptual example of a partially nested copula (right).

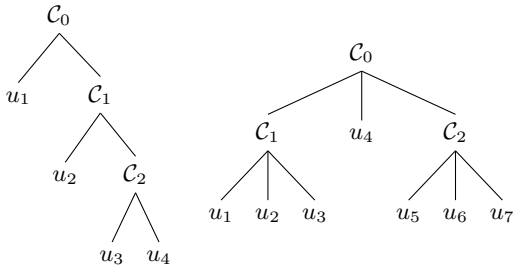

Figure 1: Examples of fully nested (left) and partially nested (right) copulas.

3.3 Data Set

In order to evaluate the use of copulas for high-dimensional relevance spaces, we use the *MSLR-WEB10K* and *WEB30K* datasets, two publicly available collections of 10,000 (30,000, respectively) real Web search queries and an annotated sample of hundreds of thousands of related impressions. For each query-url pair, a set of 136 features are available. The majority of the feature space considers dimensions related to topicality such as tf/idf scores or query term frequencies in different sections of the page. There are, however, several features that capture alternative relevance criteria such as general page authority, quality or textual complexity. For an overview of the full list of features, please consult the data set Web page (`http://research.microsoft.com/en-us/projects/mslr/feature.aspx`). The corpora are pre-partitioned into 5 equally sized folds to allow for cross validation in a 3-1-1 split of training, validation and test sets.

4. EXPERIMENTS

This section discusses our experimental set-up and findings. All experimental results were obtained by means of 5-fold cross validation on the *MSLR-WEB10K* and *MSLR-WEB10K* datasets. In order to set the performance of the various copula models into perspective, we include a common weighted linear combination scheme l as a baseline. Concrete settings of the mixture parameters λ_i are determined based on a greedy parameter sweep (ranging from $0 \ldots 1$ in steps of 0.005) on the training set of each CV fold.

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

Additionally, we compare to LambdaMART [11], a competitive learning-to-rank baseline. The relevant model parameters are tuned on the validation set. We rely on the implementation of the GBM package for R (`http://cran.r-project.org/web/packages/gbm/`).

We investigate three types of copula models. The "flattest" nesting hierarchy is given by copulas without any subnesting. All 136 dimensions are included in a single model, describing all inter-dimensional dependencies by a single parameter θ. This strategy is equivalent to the method presented in [3]. To study some simple, yet indicative examples of nested copulas, we include a fully nested approach in which the nesting order is determined randomly and the average results across 50 randomizations are reported. Note that the concrete nesting structure is an additional degree of freedom that holds significant modelling power. We leave this aspect out of the scope of this paper. Finally, as an example of partially nested copulas, we rely on the existing semantic grouping of dimensions in the dataset (e.g., all *tf/idf* features, or all *query term coverage* features) and estimate individual copulas $\mathcal{C}_{d_i}$ for each group d_i. All of these group copulas are then combined in an overall copula $C_{partial}(U)$. Groups that comprise only a single dimension are directly included into $\mathcal{C}_{partial}(U)$. This will become especially relevant for the experiments presented later in Figure 2. For the extreme case of exclusively single-dimensional groups, this model becomes equivalent to non-nested copulas. All copula experiments presented in this paper are based on the publicly available implementation for R [12].

Table 1: Performance comparison of copula and L2R models.

Method	$nDCG_{10K}$	$nDCG_{30K}$	MAP_{10K}	MAP_{30K}
Linear Comb.	0.49	0.46	0.30	0.28
LambdaMart	0.56*	0.55*	0.37*	0.37*
Copula	0.51	0.50*	0.32	0.32*
Nested	0.53*	0.54*	0.33*	0.33*
Fully Nested	0.54*	0.54*	0.36*	0.35*

For model comparison, we use two well known metrics: normalized Discounted Cumulative Gain (nDCG) and Mean Average Precision (MAP). Table 1 shows the resulting cross-validation performance of the respective methods on the full 136-dimensional datasets. Statistically significant performance improvements with respect to the linear combination baseline are denoted by the $*$ character. We used a Wilcoxon signed rank test with $\alpha < 0.05$ confidence level.

We can note a clear ordering of approaches in which linear combinations achieve the lowest overall performance and the LambdaMART method delivers the best results. The various copula-based models lie between these extremes. Global copulas show slightly better performance than a linear feature combination, these differences were, however, not found to be significant. For both forms of nested copulas, we can note significantly higher scores in terms of nDCG and MAP. With respect to our research questions, we note that copula-based models, especially nested ones, show strong ranking performance for high-dimensional settings. Fully nested copulas, especially, approximate the performance of the learning-to-rank model to the degree, that we could not note any statistically significant differences between the two methods.

In order to further investigate the individual performances of the various methods as the dimensionality of the relevance space increases, we modify the setting by varying the number of dimensions k between 1 and 136. Figure 2 shows the results of this experiment in terms of nDCG and MAP on the *MSLR-WEB10K* dataset. The figures for the *WEB30K* dataset are omitted to save space, as they display identical tendencies. For each choice of k, we randomly sample $n = 100$ feature subsets, train the respective models on each set and average the resulting retrieval performance. For all methods, we note steep performance gains with each dimension that is added early on. These improvements slowly level out and reach a largely stable performance for relevance spaces of size $75 \leq k \leq 136$. An especially noteworthy observation can be made in the comparison of global copulas and linear combination models. While early on, linear models show higher scores in both metrics, this tendency re-

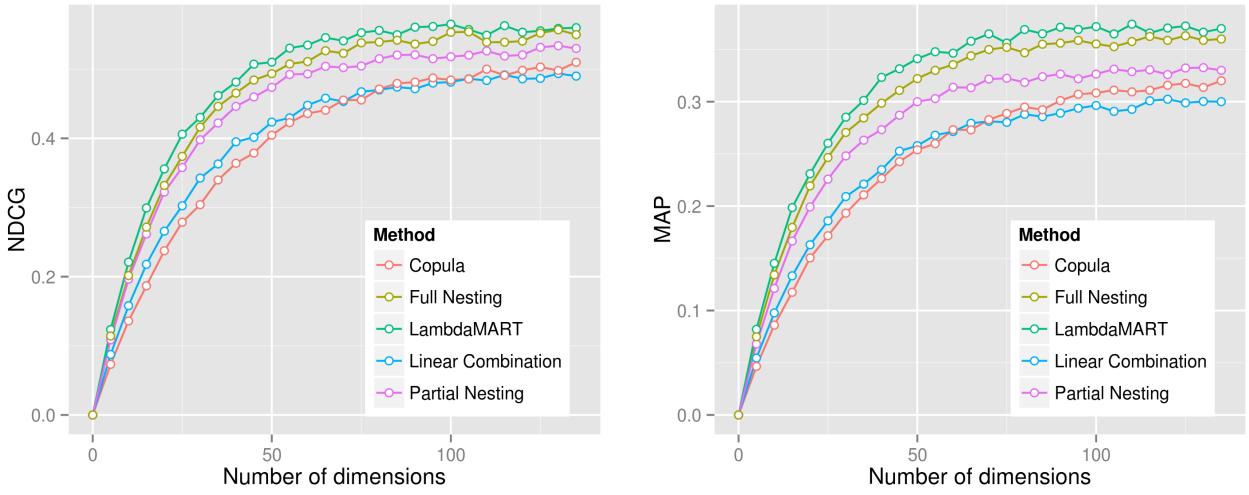

Figure 2: NDCG and MAP as functions of the dimensionality of the underlying *MSLR-WEB10K* relevance space.

verses for high-dimensional spaces ($60 \leq k \leq 80$). Previous work [3] noted competitive ranking performance of linear combination models for most of their experimental corpora. The authors concluded that the individual score distributions inherent to the respective domains may favour either of the approaches. Their study was, however, limited to two-dimensional relevance estimates. As we can see from the current example, even in domains that are seemingly well represented by linear models, copulas can achieve performance gains as the problem scales up in dimensionality.

5. CONCLUSION

This paper presents a piece of ongoing work, investigating the performance of copula-based relevance models for application in information retrieval. In particular, we looked at high-dimensional relevance spaces and used the *MSLR-WEB10K* and *WEB30K* datasets, two established learning-to-rank resources, as our experimental domain.

Our experiments suggest that for high-dimensional settings, copulas show significantly greater retrieval performance than linear combination models. When iteratively increasing the dimensionality of the relevance space, we note a widening gap between copula and linear fusion performance. Additionally, we found nested copulas to perform especially well when the number of dimensions increases.

We foresee several promising directions for future expansions of this work. In this paper, we gave an initial performance comparison of standard and nested copulas. Nested Archimedean copulas come with a high number of degrees of freedom, the tuning of which, however, exceeded the scope of this work. Here, we investigated fully nested copulas (of arbitrary nesting order) as well as partially nested ones of depth 2 with individual sub copulas per feature family. In the future, a careful investigation of nesting strategies should be conducted to find optimal order and nesting depth. In this regard, we are especially interested in applying machine learning techniques in order to construct copulas more effectively. During our investigation of relevance spaces of different dimensionality, we noted that optimal performance can be achieved long before the full pool of dimensions was included. In this paper, we randomly sampled dimensions in order to investigate the effect of dimensionality on the performance based ranking of approaches. In the future, it would be interesting to investigate active learning techniques to efficiently construct relevance spaces for copulas using the minimal subset of dimensions that results in optimal ranking performance. In this way, a significant overhead in feature extraction and model training can be made redundant.

6. REFERENCES

[1] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *ICML 2005*. ACM.

[2] Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Sebastian de la Chica, and David Sontag. Personalizing web search results by reading level. In *CIKM 2011*. ACM.

[3] Carsten Eickhoff, Arjen P. de Vries, and Kevyn Collins-Thompson. Copulas for Information Retrieval. In *SIGIR 2013*. ACM.

[4] Carsten Eickhoff, Pavel Serdyukov, and Arjen P De Vries. A combined topical/non-topical approach to identifying web sites for children. In *WSDM 2011*. ACM.

[5] Paul Embrechts, Filip Lindskog, and Alexander McNeil. Modelling dependence with copulas and applications to risk management. *Handbook of heavy tailed distributions in finance*, 2003.

[6] Shima Gerani, ChengXiang Zhai, and Fabio Crestani. Score transformation in linear combination for multi-criteria relevance ranking. In *Advances in Information Retrieval*. Springer, 2012.

[7] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR 2002*. ACM.

[8] Stefano Mizzaro. Relevance: The whole history. *JASIS*, 48(9), 1997.

[9] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Simple bm25 extension to multiple weighted fields. In *CIKM 2004*. ACM.

[10] Tefko Saracevic. Relevance reconsidered. In *CoLIS*. ACM Press, 1996.

[11] Qiang Wu, Chris J. C. Burges, Krysta M. Svore, and Jianfeng Gao. Ranking, boosting, and model adaptation. *Tecnical Report, MSR-TR-2008-109*, 2008.

[12] Jun Yan. Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, 21(4), 2007.