Ranking Related Entities for Web Search Queries

Changsung Kang, Srinivas Vadrevu, Ruiqiang Zhang, Roelof van Zwol,
Lluis Garcia Pueyo, Nicolas Torzec, Jianzhang He, Yi Chang
Yahoo! Labs
Sunnyvale, CA, USA
{ckang,svadrevu,ruiqiang,roelof,lluis,torzecn,jhe,yichang}@yahoo-inc.com

ABSTRACT

Entity ranking is a recent paradigm that refers to retrieving
and ranking related objects and entities from different struc-
tured sources in various scenarios. Entities typically have
associated categories and relationships with other entities.
In this work, we present an extensive analysis of Web-scale
entity ranking, based on machine learned ranking models
using an ensemble of pairwise preference models. Our pro-
posed system for entity ranking uses structured knowledge
bases, entity relationship graphs and user data to derive use-
ful features to facilitate semantic search with entities directly
within the learning to rank framework. The experimental
results are validated on a large-scale graph containing mil-
lions of entities and hundreds of millions of entity relation-
ships. We show that our proposed ranking solution clearly
improves a simple user behavior based ranking model.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms

Algorithms, Experimentation

Keywords

entity ranking, structured data, object ranking

1. INTRODUCTION

The availability of large collections of structured data on
the Web enables a realm of possibilities beyond the basic
textual Web search. Popular Web search engines are already
providing a rich experience, mining structured data, query
logs and web documents to provide rich information in the
search results (e.g. movie plot, genre, cast, review and show
times at the user location) or direct answers to the users
(e.g. "date of birth Brad Pitt"), while displaying related
news articles, images, videos and tweets for queries about
popular persons, organizations, media works and locations
whenever possible.

We propose to enhance this experience by providing well-
qualified related entities. In the Web search experience, the
related entities are shown as suggestions on the left-hand rail
of the search results page. The types of the related entities
to show depend upon the category of the query entity. For
example, for movie queries, the goal is to show both lateral

information in terms of related movies and faceted informa-
tion in terms of cast information. The challenge that we
propose to address in this work is to select the appropriate
set of related entities depending upon the queried entity and
its type, and to rank them in the order of relevance.

We cast the problem of entity ranking as a supervised
machine learning problem with the goal of predicting the
relevance of the related entity to the query entity. While
the previous work in this area [2] focuses on optimizing the
Click Through Rate (CTR) of the related entities alone, we
present an approach to jointly learn the relevance among the
entities using both the user click data and the editorially
assigned relevance grades.

2. FEATURE SPACE ANALYSIS

As introduced in van Zwol et al. [3], we have setup a frame-
work to uniformly compute a set of features from various
ranking sources. In this paper, we use this framework to
compute the features over Web search query logs, tags used
to annotate photos in Flickr, and entity pair detected in
tweets from Twitter users. For each source we can compute
both term based, and session based statistics. Some ex-
amples of features are entity probability, point-wise mutual
information and joint and conditional probability between
two entities and combinational features that are derived by
linear combination of simple features.

Knowledge Base
In addition to probabilistic features, we also compute graph
features that an entity network that is deduced from the
knowledge base that we build from various data sources. The
system we have designed and implemented for building this
knowledge base is called Yalinda. Yalinda extracts various
form of knowledge - including entities, their attributes (i.e.
reference, canonical name, variants, type, subtypes, other
attributes) and the relationships between them (i.e. labeled
directed binary relationships) using data scraping and data
feed processing techniques. The extracted knowledge is nor-
malized and serialized into semantic graphs, one per input
source and domain, providing unified views convenient for
consumption. Yalinda is implemented as a framework con-
sisting of general modules providing the common features
and pluggable modules providing special features such as
wrappers for specific data.

An entity graph can be constructed from this knowledge
base by simply connecting the facet of one pair to the entity
of another pair if the two have the same canonical form. The
entity network is very similar to a social network. Each node

in the social network refers to a user, while this equals an entity in the entity network. Some simple features that can be extracted from this entity network include the shortest distance between two entities, number of paths between two nodes, and the number of shared connections. The concept of shared connections is inspired from the idea of mutual friends in social networks [1]. The intuition is that if two entities have many shared nodes or connections in the entity graph, they are more related to each other.

3. MACHINE-LEARNED RANKING FOR ENTITIES

In our entity ranking problem, a set of n documents $\mathcal{D}$ is split into a set of groups by categories of results: $\mathcal{D} = \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_m$ where $\mathcal{D}_i$ contains results with category i. Our goal is to generate a ranking for each $\mathcal{D}_i$.

Pairwise Comparison Model
To generate highly accurate pairwise preferences as additional training data, we propose the *pairwise comparison model (PCM)*, which takes a "pairwise" feature vector as input and predicts a preference. Given two facets f_i and f_j, we extract a pairwise feature vector $\mathbf{w}_{ij}$ is constructed from multiple click models (e.g., SkipAbove). The training data for each training entity e is $\mathcal{T}_e = \{(\mathbf{w}_{ij}, l_i - l_j) \mid i, j \in \{1, \ldots, N\}, i \neq j\}$ where l_i is a numerical label given by human editors to facet i out of a finite set of labels L and N is the number of facets to be ranked for e. We apply the gradient boosting algorithm to obtain a function $h(\mathbf{w}_{ij})$ which predicts the relative relevance of two facets f_i and f_j. A benefit of this approach is that we can combine multiple click models through features ($\mathbf{w}_{ij}$) and thus obtain reliable preferences.

Training Ranking Function
We propose a new loss function to incorporate facet categories. We follow the gradient boosted decision trees framework applied to pairwise preferences [4]. We have two sets of pairwise preferences: $\mathcal{P}_E = \{(f_i, f_j) \mid l_i > l_j\}$ where l_i is a numerical label given by human editors to a facet f_i (the larger, the more relevant) and $\mathcal{P}_C = \{(f_i, f_j) \mid h(\mathbf{w}_{ij}) > \lambda\}$ where h is the pairwise comparison model and λ is a threshold to obtain reliable preferences.

For each (entity, facet) pair, we extract a feature vector $\mathbf{x}$ containing all the features described in Section 2. $\mathcal{P}_E$ and $\mathcal{P}_C$ can be split into two sets: $\mathcal{P}_E = \mathcal{P}_E^{inter} \cup \mathcal{P}_E^{intra}$, $\mathcal{P}_C = \mathcal{P}_C^{inter} \cup \mathcal{P}_C^{intra}$ where "*inter*" ("*intra*") denotes facets with different (the same) categories. Our loss function is

$$\frac{\alpha(1-\delta)}{|\mathcal{P}_E|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_E^{inter}} \max(0, 1 - (f(\mathbf{x}_i) - f(\mathbf{x}_j)))^2 +$$

$$\frac{1-\delta}{|\mathcal{P}_E|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_E^{intra}} \max(0, 1 - (f(\mathbf{x}_i) - f(\mathbf{x}_j)))^2 +$$

$$\frac{\alpha\delta}{|\mathcal{P}_C|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_C^{inter}} \max(0, 1 - (f(\mathbf{x}_i) - f(\mathbf{x}_j)))^2 +$$

$$\frac{\delta}{|\mathcal{P}_C|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_C^{intra}} \max(0, 1 - (f(\mathbf{x}_i) - f(\mathbf{x}_j)))^2$$

where α controls the weight for "inter"-category pairs of facets and δ is a parameter that controls the balance between $\mathcal{P}_E$ and $\mathcal{P}_C$.

Inter-category Weight	DCG-1 Gain	DCG-5 Gain	DCG-10 Gain	Pairwise Accuracy Gain
$\alpha = 0.0$	2.50%	1.97%	0.88%	20.68%
$\alpha = 0.2$	2.58%	1.95%	0.88%	20.87%
$\alpha = 0.4$	2.52%	1.98%	0.86%	20.90%
$\alpha = 0.6$	2.41%	1.97%	0.86%	20.90%
$\alpha = 0.8$	2.45%	1.94%	0.87%	20.82%
$\alpha = 1.0$	2.14%	1.96%	0.84%	20.82%

Table 1: Relevance improvements with various inter-category weights over the baseline.

4. EXPERIMENTAL RESULTS

We use the query log from Yahoo! Search Engine to sample entity queries that match from our dictionary of entity names. For each of these entities, we extract the related entities from its connections in the entity graph. Our evaluation data set consists of 6000 query entities and overall 33000 entity-facet pairs including both training and test data. These entity-facet pairs were given a five-point editorial grade that indicates the relevance of the facet entity to the query entity. Our baseline is a simple linear combination of the conditional entity probabilities, across different feature sources such as Web Search Query Terms, Web Search Query Logs, Flickr and Twitter.

Table 1 shows the DCG gains with various inter-category weights α over the baseline, which is a linear combination of the conditional probabilities across various feature data sources. The result shows that α between 0 and 1 provides the best relevance, which implies that the inter-category relationships between facets help the intra-category ranking (ranking within each group of facets with the same category).

5. CONCLUSIONS

In this paper, we presented a system for ranking related entities in the context of the Web search. We also proposed novel techniques for entity ranking based on machine learned ranking models using an ensemble of pairwise preference models. We showed how to work with categories in the context of entity ranking by introducing inter-category and intra-category weighting. We showed the results on one of the large knowledge base containing millions of entities and hundreds of millions of relationships. The experiments reveal that our proposed ranking solution clearly improves a simple user behavior based ranking model.

6. REFERENCES

[1] Y. Jin, Y. Matsuo, and M. Ishizuka. Ranking entities on the web using social network mining and ranking learning. In *WWW*, 2008.

[2] Z. Nie, Y. Zhang, J. rong Wen, and W. ying Ma. Object-level ranking: Bringing order to web objects. In *WWW*, 2005.

[3] R. van Zwol, L. G. Pueyo, M. Muralidharan, and B. Sigurbjörnsson. Ranking entity facets based on user click feedback. In *IEEE International Conference on Semantic Computing*, 2010.

[4] Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun. A general boosting method and its application to learning ranking functions for web search. In *NIPS*, pages 1697–1704. MIT Press, 2008.