2015, 15, pp. 276-288.

9. Gibson J.J A Theory of Direct Visual Perception. In: J. Royce, W. Rozenboom (eds.). The Psychology of Knowing. Gordon & Breach, New York, 1972, 270 p.

УДК 681.3.06

Оценка достоверности результатов поиска информации

Котов Эдуард Михайлович

Старший преподаватель каф. ИАСБ ИКТИБ ЮФУ

Россия, г. Таганрог, emkotov@sfedu.ru

Аннотация. Решая задачу ранжирования результатов информационного поиска, всегда возникает вопрос об эффективности функции ранжирования, насколько нам подходят традиционные меры оценки информационного поиска действительно ли исследуемый метод лучшее другого и если производилась оценка более чем для одного запроса, как все это отразится на достоверности поисковых результатов. Эти темы исследует автор статьи.

Ключевые слова: Информационный поиск, ранжирование, меры оценки.

Evaluating the reliability of information search results

Eduard Kotov

Senior Lecturer at the Department of Information and Analytical Security Systems, ICTIS, SfedU

Russia, Taganrog, emkotov@sfedu.ru

Abstract. When solving the problem of ranking information search results, there is always a question about the effectiveness of the ranking function, whether traditional measures of evaluating information search are suitable for us, whether the method under study is better than another, and if more than one query was evaluated, how all this will

affect the reliability of search results. These topics are investigated by the author of the article.

Keywords: Information search, ranking, evaluation measures.

При реализации информационного поиска возникает необходимость решения задачи ранжирования поисковых результатов, а именное проведение сортировки веб-страниц, найденных документов и т. п. в соответствии с предлагаемым поисковым запросом. Задача ранжирования заключается в сортировке набора элементов в зависимости от их степени релевантности, под которой понимается отношение элемента – документа или ссылки на документы, к некоторому объекту – поисковому запросу.

Для некоторого набора документов D_n и поискового запроса Q поисковая система оценивает все документы в наборе D согласно их предполагаемому отношению к запросу:

$$f(D, Q) = (d_1, d_2, ..., d_n)$$

Возникает вопрос: насколько эффективна функция ранжирования f?

Функция $rel(d_i)$ отражает что, $rel(d_i)=1$, если документ, стоящий на k-й позиции в ранжированном списке результатов поиска для пользователя u, релевантен. В противном случае $rel(d_i)= 0$.

Известны традиционные меры оценки информационного поиска [1].

Точность на K элементах – Precision at K (p@K):

$$P@k = \frac{1}{k}\sum_{1-i}^{k} rel(d_i).$$

Данную метрику возможно отнести к базовой метрике качества ранжирования. Применима для одного объекта. Результатом работы алгоритма ранжирования будет являться набор оценок релевантности для каждого результата поиска и отобрав среди них первые k-документов с наибольшим весом, который сопоставлен каждому поисковому результату и характеризует степень релевантности найденного документа запросу, можно посчитать долю релевантных.

Reciprocal rank (RR):

$$\mathrm{RR} = \max_{i \geq 1} \left\{ \frac{\mathrm{rel}(d_i)}{i} \right\}.$$

Метрика проста и представляет собой величину, равную обратному ранку первого правильно угаданного элемента.

Average precision (AP):

$$\mathrm{AP} = \frac{1}{R} \left(\sum_{1-i}^{n} \mathrm{rel}(d_i) * P@i \right),$$

где R – общее количество релевантных документов в коллекции.

Метрика корректирует недостаток метрики p@K, которая не учитывает порядок элементов среди лучших выбранных и равна сумме p@i только для релевантных элементов.

Если производилась оценка более чем для одного запроса, то эти меры обычно обобщаются, посредством вычисления их среднего значения: MAP (Mean average precision), MRR (Mean Reciprocal Rank):

$$MAP = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{1}{R} \left(\sum_{1-i}^{n} rel(d_i) * P@i \right) \right)_j,$$

$$MRR = \frac{1}{N} \sum_{j=1}^{N} \left(max_{i \geq 1} \left\{ \frac{rel(d_i)}{i} \right\} \right)_j.$$

Традиционная система оценка результатов информационного поиска предполагает, что для достижения полной релевантности информации должна быть доступна вся коллекции документов. Для больших наборов данных, например, в интернете – это нереально. Однако, в большинстве случаев нас интересует относительное выполнение поискового метода. Тогда справедливы вопросы: «Действительно ли метод A лучшее, чем метод B?», «Для каких запросов метод C работает хуже, чем метод D?».

Подход TREC (Text Retrieval Conference – серия конференций, исследования информационного поиска, его различных областей и задач) к сравнительной оценке сводится к следующему:

– Объединение: пусть несколько систем представляют свои рейтинги для набора поисковых запросов.

– Для каждого запроса выбрать лучшие p документы для ранжирования (как правило, $50 \leq p \leq 100$) и оценивать их по релевантности.

Такой подход приводит к неполному набору суждений о релевантности (в терминологии TREC "Qrels"). Значения метрики MAP будут неверными. Однако, каждая система вносит одинаковое количество документов, отсюда возникает равная дискриминация в отношении любой системы, и оценочные меры, такие как P@20 и P@50 будут корректны, вследствие чего возможно предположить, что взаимный ранг скорее всего будет правильный.

Представляется целесообразным использовать существующие наборы суждений о релевантности (TREC Qrels) для оценки новых метрик качества ранжирования, даже если они не участвует в TREC. Причем, отметим, что Qrels не до конца объективны: для новой методики отсутствует возможность внести свой вклад в пул и остается неопределенным, как неоцененный документ должен повлиять на оценку вычисления – предполагать, что он нерелевантен или игнорировать, и насколько надежны результаты, полученные таким образом.

Возможна реализация подхода, игнорирующего при оценке ранее неоцененные документы [2].

bpref (Buckley&Voorhees):

$$1 - \sum_{r \in R} \frac{\left|\{n \in N_{|R|}\}\right|}{|R| * \min\{|R|, |N|\}},$$

RankEff (Gronqvist):

$$1 - \sum_{r \in R} \frac{|\{n \in N\}|}{|R| * |N|},$$

P@k(j) (ad-hoc measure):

$$\frac{1}{k}\sum_{j\in J_k} rel(j).$$

R – множество оцененных релевантных документов,

N – множество оцененных нерелевантных документов,

$N_{|R|}$ – наилучшие |R| оцененные нерелевантные документы при ранжировании,

j_k – k-наилучших оцененных релевантных документов.

С другой стороны, имеет место подход, использующий алгоритмы классификации текста, чтобы устранить смещение от Qrels. Предположим, что пул был создан путем отбора топ-50 документов с каждого прогона. Если обучить классификатор документов учитывать существующие суждения Qrels, то возможно предсказать релевантность всего недооцененного набора документов из коллекции топ-50.

Рассмотрим возможности применения данного подхода на следующих методах классификации:

- SVMlight – реализация поддерживающей векторной машины;
- расхождение Кульбака-Лейблера (KLD).

SVM – это линейный классификатор, который вычисляет решение основываясь на функции:

$$f(\vec{x}) = sign(\vec{w}^T * \vec{x} + b)$$

Документы были преобразованы в векторы признаков в векторном пространстве, ограничивающем документы наиболее часто встречающимися терминами, и применяются TF-IDF правила. Предполагалось, что каждый неоцененный документ с f(x)>0 является релевантным и все остальные документы не релевантны.

Дивергенция Кульбака-Лейблера (KLD) являющийся несимметричной мерой удалённости друг от друга двух вероятностных распределений P и Q (например, языковая модель юниграммы):

$$KLD(P,Q) = \sum_x Pr\lfloor x|P\rfloor * log\left(\frac{Pr\lfloor x|P\rfloor}{Pr\lfloor x|Q\rfloor}\right),$$

Q – юниграммы, построенные из конкатенации всех оцененных релевантных документов;

P – юниграммы языковой модели, построенные на основе документа, релевантность которого должна быть определена.

Если KLD (P, Q) > Θ, то документ принимаем за релевантный. В противном случае – нет. Выбираем Θ так, чтобы точность (precision) и полнота (recall) классификатора были равны на исследуемом наборе данных: precision=recall.

Оценивая оба подхода с использованием существующих данных TREC, можно резюмировать следующее:

– на практике возможна оценка нового метода с использованием существующего суждения о релевантности;

– традиционные меры (P@k, AP, MRR) обычно недооценивают эффективность новых методов;

– алгоритм классификации текста (SVM) может быть использован для прогнозирования релевантности для неоцененных документов;

– подходящим критерием оптимизации при обучении классификатора является precision=recall.

Библиографический список

1. Yue Shia, Alexandros Karatzogloub, Linas Baltrunas. CLiMF: learning to maximize reciprocal rank with collaborative less-is-more filtering. RecSys’12 the sixth ACM conference on Recommender systems, pages 139-146, 2012.

2. Qiang Wu, Christopher J. C. Burges, Krysta M. Svore. Adapting Boosting for Information Retrieval Measures. Information Retrieval, Vol. 13, pages 254 - 270, 2010.