Инженерный вестник Дона, №1 (2019)
ivdon.ru/ru/magazine/archive/n1y2019/5660

Разработка алгоритма поиска и ранжирования неструктурированной информации

А.И. Квач, А.В. Высочкин, Е.М. Портнов

Национальный исследовательский университет «Московский Институт Электронной Техники»

Аннотация: Проведен анализ используемых алгоритмов поиска неструктурированной информации, в ходе которого установлено, что в них не учитывается тематическая ориентированность специализированных коллекции документов. Предлагается использовать алгоритм имитации отжига для получения численных коэффициентов используемых в алгоритме ранжирования информационно-поисковой системы. Получены результаты работы стандартного и модифицированного алгоритма поиска и ранжирования результатов поиска неструктурированной информации. В результате значение метрик качества удалось улучшить в среднем на 8%, а среднее значение ошибки снизилось на 29%.

Ключевые слова: информационный поиск, ранжирование, модель, метрика качества, релевантность, неструктурированная информация.

Важнейшим показателем оценки качества информационно-поисковых систем (ИПС) является степень удовлетворенности пользователя, которая зависит от следующих факторов: скорость поиска информации и соответствие результатов работы системы поисковому намерению ее пользователя (релевантность поиска) [1-6]. Оценки качества работы ИПС производится с помощью различных метрик, при этом, алгоритм, для которого оценка совпадает с оценкой асессоров, считается более приоритетным [3,7]. Ниже приведен обзор алгоритмов расчета метрик оценки качества работы ИПС.

Для оценки качества поисковых систем по первым *n* документам применяются *DCG@n* и *NDCG@n*. Данные метрики являются одними из первых, которые широко распространены graded-метрик.

Модификации данных метрик использовались на РОМИП для оценки качества поиска [8]:

$$DCG@n = \sum_{p=1}^{n} \frac{2^{grade(p)}-1}{\log_2(2+p)}, \quad (1)$$

$$NDCG@n = \frac{DCG@n}{Z}, \quad (2)$$

grade(*p*) – значение средней оценки релевантности присвоенной экспертамидля *p*-го документа в коллекции результатов выполнения запроса (*grade*$\in$[0,3]); $\frac{1}{log2(2+p)}$ – дисконтное значение, присваиваемое определенную позицию документа в коллекции результата поиска;

Z – фактора нормализации, соответствующий максимально возможному значению *DCG@n* для поискового запроса.

Исходя из этого: *NDCG* $\in$ [0,1] , а *NDCG*=1 только в случае убывающего ранжирования по экспертной оценки для документов результатапоиска.

Согласно представленной модели, степень влияния оценки значения релевантности документа коллекции на конкретную метрику зависит как от позиции документа в результате выполнения поискового запроса, так и отчисленной оценки значений релевантности документов, которые находятся выше в списке результатов [3,5,9].

Метрика оценки качества GradedMeanReciprocalRank использовалась как основная метрики на конференции YahooLearningtoRankChallenge и находилась по следующей формуле:

$$ERR = \sum_r \frac{1}{r} - p(userstopatposition r), E = mc^2, \quad (3)$$

$$p(userstopatposition r) = R_r \prod_{i=1}^{r-1}(1 - R_i), \quad (4)$$

$$R_i = \frac{2^{grade(i)}-1}{2^{max_{grade}}}. \quad (5)$$

В социальных науках одним из самых распространённым показателем согласованности численных оценок есть *каппа-статистика (kappa statistics)*, которая была разработана для «категориальных» оценок, и она учитываем поправку навозможное совпадение оценок:

$$Kappa = \frac{P(A)-P(E)}{1-P(E)}. \quad (6)$$

где параметр $P(A)$— значение доли совпавших оценок асессоров; параметр $P(E)$ —значение ожидаемой доли возможно совпавших оценок асессоров.

Численное значение каппа, которое превышающее 0,8 – согласование хорошее, если от 0,67 и до 0,8 — согласование удовлетворительное, а еслименьше 0,67 — то согласование сомнительное. Однако точные численныезначения порогов зависят от предназначения коллекции данных.

Обычно, значение каппа-статистика находится в среднем (удовлетворительном) диапазоне данных. При этом бинарные выводы асессоров о значениях релевантности документов очень слабо согласованы, данная ситуация является причиной, по которой более подробная шкала не используется [10].

Преимущество численной оценки поисковой системы на основе стандартной модели релевантных/нерелевантных результатов заключается в конкретных условиях, в которых следует проводить сравнение результатов экспериментов с участием разных ИПС или модификаций одной из поисковой системы. Такой способ формального тестирования системы значительно дешевле, а также позволяет более подробно выявлять влияние значений параметров поисковой системы на качество её работы.

У методов оценки качества работы поисковой системы на основезначений релевантности документов существует проблема разницы между значениями релевантности и значениями маргинальной релевантности. Математические модели ИПС не знают, сохраняет ли документ значение своей полезности, после просмотра пользователем других документов [10]. Документы с высоким показателем релевантности соответствует потребностям конечного пользователя, но также они могут содержать вторичные данные, которые также могут содержаться в других источниках.

Довольно часто используемые сочетание метрик для оценки качества работы систем: 1) для информационного поиска: значение полноты (recall); значение точности (*precision*); значение аккуратности (*accuracy*); значение ошибки (*error*); значение F-меры (*F-measure*). 2) для аннотирования по поисковому запросу: значение точности (*precision*); значение аккуратности (*accuracy*); значение ошибки (*error*); 3) в вопросно-ответном информационном поиске: значение точности (precision); значение усредненной ценности ответов поискового запроса (TrecReciprocalRank); значение усредненной ценности ответов поискового запроса(RomipReciprocalRank). Для многих метрик оценки качества работы ИПС является неопределенной ситуация, когда для данного поискового запроса не нашлось релевантных документов.

В результате анализ используемых поисковых алгоритмов выявленявный недостаток: в используемых алгоритмах не учитывается тематическая ориентированность специализированных коллекции документов. Данная проблема актуальна в различных сферах и бизнес-процессах, в которых очень сильно отличаются типы документов, а каждый рассматриваемый тип важно ранжировать по своим определенным признакам. Задачи такого типа принадлежат к классу NP-сложных задач. Предлагается использовать алгоритм имитации отжига для получения численных коэффициентов, используемых в алгоритме ранжирования информационно-поисковойсистемы.

Алгоритм имитации отжига основывается на физической модели процесса кристаллизации вещества, происходящей при отжиге металлов. Процесс кристаллизации протекает при снижающейся температуре, которая начинает убывать с некоторого начального значения. В процессе кристаллизации, атомы постепенно выстраиваются в кристаллическую решетку и, на протяжении некоторого времени, обладают энергетическим потенциалом, достаточным для осуществления перехода между узлами

кристаллической решетки. На начальном этапе процесса, когда температура еще достаточно высока, вероятность перехода атомов больше. По мере снижения температуры, происходит уменьшения энергетического потенциала и вероятность перехода становится меньше, чем в начале процесса.

Классический алгоритм имитации отжига описывается следующим образом: 1. Запустить процесс из начальной точки x при заданной начальной температуре $T = T_{max}$ 2. Пока $T > 0$ повторить K раз следующие действия:
• выбрать новое решение $x`$ из окрестности и $x$; • рассчитать изменение целевой функции $\Delta = E(x) — E(x`)$; • если $\Delta < 0$ принять $x \leftarrow x`$ в противном случае (при $\Delta <= 0$)принять, что $x \leftarrow x$с вероятностью $exp(-\Delta/T)$ путем генерации случайного числа R из интервала $(0,1)$ с последующим сравнением его со значением $exp(-\Delta/T)$;если $exp(-\Delta/T) > R$ принять новое решение; в противном случае проигнорировать его. 3. Уменьшить температуру $T \leftarrow rT$с использованием коэффициента уменьшения r выбираемого из интервала $(0,1)$, и вернуться к шагу 2. В методе имитации отжига, вероятность перехода к новому значению вычисляется в соответствии с распределением Гиббса (8)

$$P(X^* \to X_{t+1}|X_t) = \begin{cases} 1, E(X^*) - E(X_t) < 0 \\ \exp\left(-\frac{E(X^*)-E(X_t)}{T_t}\right), E(X^*) - E(X_t) \geq 0 \end{cases} . \quad (8)$$

Варьируя закон убывания температуры T, можно ускорить процесс протекания имитации отжига. Вообще говоря, температура может быть представлена положительными элементами убывающей последовательности.

Далее рассмотрим этапы работы разработанного алгоритма.

Этап 1. Осуществляется загрузка коллекции документов из гетерогенных источников данных (FTP, DB, FS). Производится проверка времени изменения документа. Для новых или измененных документов производится добавления данных в поисковый индекс. Для новой коллекции. производится процедура полного обновления поискового индекса.

Этап 2. Каждый документ проходит через цепочку фильтрующих функций для обеспечения очистки документа от информационного шума. Далее, каждый документ приводится к нормальной форме с помощью процедуры нормализации (процесс стемминга), при этом, основа слова не обязательно должна совпадать с морфологическим корнем слова. Данный процесс также необходим для сокращения поискового индекса и способствует увеличению скорости работы с ним.

Этап 3. Для значимых коэффициентов k1 и b случайным образом выбирается некоторые значения. Эти значения используются в методе имитации отжига в качестве начального приближения. На данном этапе происходит итерационное получения новых численных показателей для значимых коэффициентов в поисковой системе. Результаты работы поисковой системы оцениваются со значениями асессоров, до тех пор, пока значения ошибке не будет меньше, чем заданное или не завершится алгоритм имитации.

Проведение оценки разработанного алгоритма проводилось в поисковой системе Elasticsearch. В качестве запросов использовались следующие типы: информационный; транзакционный; навигационный; однословный; двухсловный; многословный. Информационный тип запросов характерен для поиска общей информации и не подразумевает при этом тематическую ориентированность запрашиваемого ресурса. Транзакционный тип запросов характерен для поиска определенной категории информации. Навигационный тип запросов характерен для поиска определенного информационного ресурса. Результаты экспериментального исследования полноты представлены в таблице 1.

Таблица 1

Сравнение значений полноты для стандартного и модифицированного алгоритма ранжирования

Типы поискового запроса	Значение метрики Recall	
	Модифицированный алгоритм ранжирования	Стандартный алгоритм ранжирования
Информационный	0.80	0,9
Транзакционный	0.71	0.83
Навигационный	0.75	0.85
Однословный	0.78	0.80
Двухсловный	0.60	0.71
Многословный	0.70	0.78

Среднее значение полноты поиска модифицированного алгоритма спомощью метода имитации отжига:

$$R_m^* = \frac{1}{N}\sum_{i=1}^{N} R_i^* = 0{,}89.$$

Усредненное значение разницы прироста полноты, для группы типов рассмотренных поисковых запросов, составляет:

$$\Delta_R = \frac{1}{N}\left(\sum_{i=1}^{N} R_i^* - \sum_{i=1}^{N} R_i\right) = 0{,}09.$$

Основываясь на среднем значении метрики полноты, прирост качества поиска составил ~ 11% .

Результаты экспериментального исследования F-меры для различныхтипов запросов представлены в таблице 2.

Таблица 2

Сравнение значений F-меры для стандартного и модифицированного алгоритма ранжирования

Типы поискового запроса	Значение метрики Recall	
	Модифицированный алгоритм ранжирования	Стандартный алгоритм ранжирования
Информационный	0.55	0.70

Транзакционный	0.48	0.62
Навигационный	0.51	0.63
Однословный	0.53	0.64
Двухсловный	0.48	0.61
Многословный	0.50	0.66

Основываясь на среднем значении метрики F-меры, прирост качества поиска составил ~ 21%.

Работа выполнялась при финансовой поддержке РФФИ (договор № 18-07-00079\18).

Заключение

Разработана математическая модель поиска неструктурированной информации и ранжирования документов в информационно-поисковой системе на основе метода имитации отжига для подбора числовых коэффициентов функции ранжирования.

Получены результаты работы стандартного и модифицированного алгоритма поиска и ранжирования результатов поиска полнотекстовой поисковой системы. В результате значение метрик качества удалось улучшить в среднем на 8%, а среднее значение ошибки снизилось на 29%.

Литература

1. А.А. Харламов, Т.В. Ермоленко, А.А. Жонин Моделирование динамики процессов на основе анализа последовательности текстовых выборок. Инженерный вестник Дона, 2013, №4. – URL: ivdon.ru/ru/magazine/archive/n4y2013/2047.

2. С.П. Воробьев, М.Б. Хорошко Модификация модели векторного пространства для ранжирования документов Инженерный вестник Дона, 2012, №3. – URL: ivdon.ru/ru/magazine/archive/n3y2012/976/.

3. Портнов Е.М., Со Тант. Формализация задачи полнотекстового поиска информации в структурированных базах знаний // Естественные и

технические науки, 2008, №3. С. 210-212.

4. Mohammad, M., Kosaraju, S., Bayramoglu, T., Modgil, G., Kang, M. Automatic knowledge extraction from OCR documents using hierarchical document analysis // Proceedings of the 2018 Research in Adaptive and Convergent Systems, RACS 2018.- pp. 189-194

5. Баин А.М., Слюсарь В.В., Со Тант. Методика автоматизированного анализа документированной информации в системах поддержки принятия решений // Известия высших учебных заведений. Электроника. – 2008. - №3. С. 128-131.

6. Портнов Е.М., Баин А.М., Чжи Я Аунг. Разработка иерархической многомодульной базы знаний с динамически управляемой структурой // Оборонный комплекс - научно-техническому прогрессу России. 2009. № 4. С. 76-80.

7. Портнов Е.М., Ломакин А.А., Стефаненко Л.Ю. Структуры селективного управления нестационарными логистическими потоками // Оборонный комплекс - научно-техническому прогрессу России. 2010. № 1. С. 57-61.

8. Агеев М., Кураленок И., Некрестьянов И., Официальные метрики РОМИП // 2010. С. 172-187.

9. Квач А.И., Портнов Е.М. Методика балансировки нагрузки в системах управления IP-шлюзами // Естественные и технические науки. 2018. № 7 (121). С. 157-159.

10. Patrick P., Vishnu V. a Joint Information Model for n-best Ranking. // COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1 2017. pp 681-688.

References

1. A.A. Kharlamov, T.V. Ermolenko, A.A. Zhonin Inženernyj vestnik Dona (Rus), 2013, №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2047.

2. S.P. Vorob'ev, M.B. Khoroshko Inženernyj vestnik Dona (Rus), 2012, №3. URL: ivdon.ru/ru/magazine/archive/n3y2012/976/.

3. Portnov E.M., So Tant.. Estestvennye i tekhnicheskie nauki, 2008, №3. pp. 210-212.

4. Mohammad, M., Kosaraju, S., Bayramoglu, T., Modgil, G., Kang, M. Proceedings of the 2018 Research in Adaptive and Convergent Systems, RACS 2018. pp. 189-194

5. Bain A.M., Slyusar' V.V., So Tant. Izvestiya vysshikh uchebnykh zavedeniy. Elektronika. 2008. №3. pp. 128-131.

6. Portnov E.M., Bain A.M., Chzhi Ya Aung. Oboronnyy kompleks - nauchno-tekhnicheskomu progressu Rossii. 2009. № 4. pp. 76-80.

7. Portnov E.M., Lomakin A.A., Stefanenko L.Yu. Oboronnyy kompleks - nauchno-tekhnicheskomu progressu Rossii. 2010. № 1. pp. 57-61.

8. Ageev M., Kuralenok I., Nekrest'yanov I., Ofitsial'nye metriki ROMIP. 2010. pp. 172-187.

9. Kvach A.I., Portnov E.M. Estestvennye i tekhnicheskie nauki. 2018. № 7 (121). pp. 157-159.

10. Patrick P., Vishnu V. COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1 2017. Pp. 681-688.