УДК 004.62

С.А. Сакулин, А.Н. Алфимцев

РАЗВИТИЕ МЕТОДА ОПРЕДЕЛЕНИЯ ВЕСОВ ДЛЯ ВЗВЕШЕННОГО ЗОННОГО РАНЖИРОВАНИЯ В ИНФОРМАЦИОННОМ ПОИСКЕ

Информационный поиск на основе взвешенного зонного ранжирования подразумевает присвоение каждой зоне или полю в метаданных документов весового коэффициента с использованием методов машинного обучения. Рассмотрен метод определения весов, в котором для вычисления взвешенной зонной релевантности вместо средневзвешенного оператора применен нечеткий интеграл Шоке. Это позволяет учесть при расчетах релевантности возможные взаимозависимости между зонными показателями, что в конечном итоге повысит точность ранжирования.

E-mail: sakulin@bmstu.ru

Ключевые слова: *информационный поиск, взвешенное зонное ранжирование, машинное обучение, оператор агрегирования, нечеткая мера, нечеткий интеграл Шоке.*

Введение. Информационный поиск представляет собой поиск документов по запросу с помощью тех или иных методов [1]. При работе с огромным числом документов результатом поиска станет настолько большое число документов, релевантных запросу, что пользователь будет не в состоянии их просмотреть. Поэтому одной из важных задач информационного поиска является ранжирование результатов по степени их соответствия запросу.

Если при таком ранжировании предполагается использование метаданных документов, то необходимо учитывать экспертные знания о структуре и особенностях этих метаданных. Здесь под метаданными понимают поля (например, дата создания документа, вид документа, стоимость книги и т. п.) и зоны (название, автор, издательство, аннотация, ключевые слова, текст и т. п.). Отличие между зонами и полями заключается в том, что поле может иметь ограниченный, заранее определенный набор значений, а на зону такие ограничения не распространяются. Далее для краткости поля и зоны будем называть зонами. В работе [1] описан метод определения весов с помощью машинного обучения на основе обучающих примеров, в котором каждой отдельной h-й зоне присваивается весовой коэффициент g_h. Произвольный текстовый запрос обозначим q, а документ — d. Каждой паре $(q,\ d)$ при взвешенном зонном ранжировании присваивается значение релевантности на единичном отрезке путем вычисления линейной комбинации зонных показателей. В эту комбинацию каждая зона вносит значение из единичного отрезка. Рассмотрим множество документов, каждый из которых имеет H зон. Пусть

$g_h \in [0,\ 1],\ 1 \le h \le H$, при этом $\sum_{h=1}^{H} g_h = 1,\ s_h \in [0,\ 1]$ — степень соответствия (или несоответствия) между запросом q и h-й зоной. Величина s_h может вычисляться по-разному для каждой из зон [1]. Наиболее простой способ ее определения может быть таким: если все термины запроса принадлежат конкретной зоне, то ее значение равно единице; если только один термин принадлежит зоне, то — $1/r$, где r — число терминов в запросе; если ни одного — нулю. В других способах определения значения s_i может использоваться частота, с которой встречается в той или иной зоне термин запроса. Способы могут быть основаны на показателях качества документа, его возрасте, длине и т. п. В частности, существует способ определения зонных показателей с помощью функции BM25F [2], которая позволяет учитывать частоты вхождения термина запроса в зоны документа. В свою очередь, эта функция основана на функции BM25 [3], представляющей собой линейную комбинацию трех ключевых атрибутов: частоты термина, частоты документа и длины документа. В данной статье особое внимание уделено агрегированию зонных показателей в единый показатель релевантности документа запросу. Для любого из способов определения зонных показателей агрегирование осуществляется путем их линейной комбинации. Таким образом, взвешенная зонная релевантность рассчитывается по формуле

$$\sum_{h=1}^{H} g_h s_h. \qquad (1)$$

Предположим, что есть множество обучающих примеров, каждый из которых является кортежем, состоящим из запроса q, документа d и оценки релевантности q и d. Обычно для любого обучающего запроса q имеется совокупность документов, полностью упорядоченная экспертом по релевантности запросу. В соответствии с таким порядком оценки релевантности могут выставляться экспертом на единичном отрезке. Тогда весовые коэффициенты g_h определяются путем машинного обучения на доступных примерах так, чтобы полученные значения коэффициентов позволяли аппроксимировать оценки релевантности из обучающих примеров. Расчет весовых коэффициентов сводится к задаче оптимизации, целевая функция которой представляет собой суммарную ошибку, соответствующую множеству обучающих примеров. Существуют эмпирические правила присвоения весовых коэффициентов зонам документа. Например, достичь высокой точности ранжирования можно, назначив относительно высокий весовой коэффициент заголовку документа, либо точность ранжирования новостных документов при поиске по запросу можно увеличить выделением первого предложения в отдельную

зону и присваиванием этой зоне повышенного весового коэффициента [4, 5]. Эти и подобные им правила могут применяться при машинном обучении в рамках средневзвешенного агрегирования зонных показателей [6].

Описанный выше способ во всех его разновидностях предусматривает неявное предположение о взаимной независимости величин s_i. Однако на основе относительно простых рассуждений можно показать, что величины s_h могут быть зависимы друг от друга. Например, если термин запроса встретился в названии новостного документа, то этот же термин встретится и в первом предложении. В этом случае имеют дело с положительной корреляцией степеней соответствия s_i и, вычисляя релевантность по формуле (1), заведомо получают некоторую избыточность результата. Это явление положительной корреляции агрегируемых величин и способа компенсации соответствующей избыточности результата подробно рассмотрено, например, в работе [7]. Рассмотрим более сложную зависимость. Пусть, по мнению эксперта, термин запроса встречается и в теле документа, и в аннотации. Тогда документ будет более релевантным запросу, если этот же самый термин содержится в поле «название документа», а не в поле «вид документа». Такая зависимость степеней соответствия s_i известна как предпочтительная зависимость критериев, которая не может быть выражена ни одним из аддитивных операторов, в том числе и средневзвешенным [7]. Подобные рассуждения эксперта невозможно формализовать в виде правил для получения весовых коэффициентов зон с помощью машинного обучения в рамках средневзвешенных операторов агрегирования. Следовательно, при применении для вычисления релевантности документов запросу средневзвешенного оператора в некоторой степени огрубляется результат с учетом предположения о независимости величин s_h друг от друга.

Нечеткие меры и интеграл Шоке. Альтернативой средневзвешенному оператору может стать интеграл Шоке по нечеткой мере. Он является обобщением средневзвешенного оператора для случая, когда величины s_h (следуя устоявшейся терминологии, будем называть критериями агрегирования) могут зависеть друг от друга [7, 8].

Нечеткая (дискретная) мера — функция множества $\psi : 2^J \to [0, 1]$ ($2J$ — множество всех подмножеств множества индексов критериев $J = \{1, \ldots, H\}$), которая удовлетворяет следующим условиям:

1) $\psi(\varnothing) = 0,\ \psi(J) = 1;$

2) $\forall D,\ B \subseteq J : D \subseteq B \Rightarrow \psi(D) \le \psi(B).$

Функция также выражает относительный вес или важность не только каждого отдельного критерия, но и всякого подмножества критериев [7]. Нечеткий интеграл Шоке [9], введенный в 1974 г. М. Суджено на основе неаддитивных мер Шоке [10], используется в качестве оператора агрегирования, позволяющего отражать знания

эксперта о зависимостях критериев путем выбора значений соответствующих параметров. Использование интеграла для построения операторов агрегирования зависимых критериев рассмотрено в работах [7, 8]. В частности, предпочтительная независимость критериев, моделируемая с помощью интеграла Шоке, изложена в работе [8]. В работе [11] проведен подробный анализ применения относительно нового метода машинного обучения на основе интеграла Шоке в различных прикладных областях и сделан вывод о целесообразности его использования. В контексте информационного поиска интеграл Шоке можно применить для моделирования экспертных предпочтений, формализованных в виде правил, аналогичных правилам, рассмотренным выше.

Нечеткий (дискретный) интеграл Шоке от критериев $s_1, ..., s_H$ по нечеткой мере ψ определяется по выражению

$$CH_{\psi}(s_1, ..., s_H) := \sum_{h=1}^{H} s_{(h)} \left[\psi(A_{(h)}) - \psi(A_{(h+1)}) \right],$$

где $(\cdot)$ означает перестановку индексов в множество J такое, что $x_{(1)} \le ... \le x_{(H)}$; $A_{(h)} = \{(h), ..., (H)\}$ и $A_{(H+1)} = \varnothing$ [8].

Идентификация нечеткой меры при взвешенном зонном ранжировании. В случае использования средневзвешенного оператора весовые коэффициенты g_h могут быть напрямую заданы экспертом. Но вследствие большой трудоемкости такого задания в подавляющем большинстве случаев они определяются на основе машинного обучения [1]. Если вместо средневзвешенного оператора применяется интеграл Шоке, то вместо весовых коэффициентов g_h требуется получить нечеткую меру ψ. Задание нечеткой меры экспертным путем еще более затруднено, чем задание весовых коэффициентов g_h вследствие экспоненциально возрастающей сложности. Так, для четырех критериев от эксперта потребуется задать $2^4 = 16$ коэффициентов нечеткой меры, что практически невозможно. Поэтому коэффициенты меры ψ определяются методами машинного обучения. Для этого необходимо сформировать множество обучающих примеров и множество формализованных эмпирических правил наподобие тех, которые описаны выше. Каждый из этих примеров является тройкой вида $\Phi_k = (d_k,\ q_k,\ r(d_k,\ q_k))$, в которой документу d_k и запросу q_k экспертным путем ставится в соответствие оценка релевантности $r(d_k,\ q_k)$ на единичном отрезке, либо эти оценки ранжируются экспертом. Правила будут представлять собой ограничения, накладываемые на нечеткую меру и на интеграл Шоке в виде нестрогих частичных порядков на множестве реализаций зонных показателей, результатов агрегирования (итоговой релевантности документа), а также индексов Шепли, критериев и индексов взаимодей-

ствия критериев. Методы формализации подобных правил подробно рассмотрены в работе [12]. В частности, правило о коррелированности зонных показателей формализуется присвоением положительного знака индексу взаимодействия этих показателей. На практике, чтобы эксперт смог сформулировать правила, часто ограничиваются рассмотрением нечетких мер порядка и интеграла Шоке второго порядка [13]. Оставаясь достаточно простым, он позволяет моделировать взаимодействия критериев, которые описываются правилами, аналогичными приведенным выше. При этом не рассматриваются зависимости между более чем двумя критериями. Для каждого обучающего примера имеются значения s_h соответствия запроса каждой зоне документа. Релевантность документа d_k запросу q_k будет определяться как $\text{score}(d_k,\ q_k) = CH_\psi(s_1,\ ...,\ s_H)$. При таком характере доступной информации можно применить метод идентификации нечеткой меры, который основан на минимизации дисперсии [12, 14]. Его достоинством является отсутствие каких-либо жестких требований к входным данным в отличие от других методов идентификации [12]. Метод реализован на принципе максимальной энтропии, который заключается в выборе нечеткой меры, учитывающей всю доступную информацию (в виде обучающих примеров и правил). Однако к недостающей информации относятся наименее предвзято, т. е. максимизируют ее неопределенность. При взвешенном зонном ранжировании документов будем придерживаться этого принципа. Целевая функция этого метода определяется как дисперсия нечеткой меры [14]:

$$F_{MV}(\psi) := \frac{1}{|J|}\sum_{i\in J}\sum_{G\subseteq J-i}\frac{(|J|-|G|-1)!|G|!}{|J|!}\left(\sum_{D\subseteq G}a(D\cup i)-\frac{1}{|J|}\right)^2.$$

Соответствующая задача оптимизации принимает следующую форму. Минимизировать меру $F_{MV}(\psi)$ при ограничениях:

$$\sum_{\substack{D\subseteq G\\ |D|\le\kappa-1}} a(D\cup i)\ge 0,\ \ \forall i\in J,\ \ \forall G\subseteq J-i;$$

$$\sum_{\substack{D\subseteq G\\ |D|\le\kappa-1}} a(D)=1;$$

$$CH_\psi(s_1,\ ...,\ s_H)-CH_\psi(s'_1,\ ...,\ s'_H)\ge\delta_{CH};$$

....

Здесь $i\in J$; $G\subseteq J$; $\kappa-$ порядок нечеткой меры ψ; δ_{CH} — задаваемый экспертом порог безразличия для сравнения двух результатов

агрегирования по Шоке; $a(D)$ — функция множества на множестве J, которая в комбинаторике называется функцией Мёбиуса по ψ, $a(D) = \sum_{G \subseteq D} (-1)^{|D|-|G|} \psi(D)$; $s_1, ..., s_H$ и $s'_1, ..., s'_H$ — значения зонных показателей для двух документов, первый из которых более релевантен запросу по мнению эксперта (отношение строгого порядка между двумя оценками релевантности).

Процедура определения весовых коэффициентов для взвешенного зонного ранжирования. При использовании в качестве оператора агрегирования интеграла Шоке по нечеткой мере эта процедура состоит из следующих шагов.

Шаг 1. Сформировать множество зон типового документа, а также способы определения зонных показателей.

Шаг 2. Сформировать в рамках коллекции документов обучающие примеры $\Phi_1, ..., \Phi_K$ в виде заданных экспертом оценок релевантности $r(d_k, q_k)$ и (или) нестрогого частичного порядка на множестве этих оценок, т. е. осуществить экспертное ранжирование документов по отношению к запросу, а также правила в виде частичных нестрогих порядков на множествах параметров интеграла Шоке.

Шаг 3. Формализовать полученную на шаге 2 информацию в виде ограничений на параметры интеграла Шоке в виде неравенств с порогами безразличия. Задать значения порогов безразличия экспертно исходя из характера обучающих примеров и применяемых шкал.

Шаг 4. Идентифицировать нечеткую меру на основе информации, полученной на шаге 3 с помощью метода минимизации дисперсии.

При добавлении к множеству обучающих примеров и множеству правил новой доступной информации процедура повторяется, начиная с шага 3. Интеграл Шоке по нечеткой мере ψ, полученной в результате этой процедуры, является оператором агрегирования зонных показателей, с помощью которого осуществляется ранжирование документов по степени их релевантности запросу.

Экспериментальное исследование. В ходе экспериментального исследования не ставилась задача создания полноценной поисковой системы. Цель исследования заключалась в получении ответа на вопрос о практической применимости относительно нового аппарата нечетких мер и интеграла Шоке в области информационного поиска.

Множество обучающих примеров включало в себя 100 терминов и около 300 документов (публикации в научных журналах). Рассматривались четыре зоны документов: заглавие, аннотация, основной текст и список литературы. Зонные показатели s_i вычислялись на основе функции BM25F [2]. Кроме того, исходными данными для машинного обучения были два эмпирических правила, аналогичные

правилам, приведенным выше. Обучающие примеры и правила являлись ограничениями, накладываемыми на интеграл Шоке и на его параметры в процессе идентификации нечетких мер. Релевантность документа запросу оценивалась по пятибалльной шкале, представляющей собой множество {0, 1, 2, 3, 4}, аналогично тому, как это выполнено в работе [6]. В этом множестве «0» означает полное несоответствие документа запросу (отсутствие релевантности), «4» — полное соответствие (документ релевантен запросу), остальные значения — это промежуточные градации релевантности. Для идентификации нечеткой меры методом минимизации дисперсии был использован специализированный свободно распространяемый пакет Kappalab [14]. Важным вопросом, возникшим в процессе идентификации, стала необходимость экспертного назначения значений порогов безразличия [15]. Эти значения выбирались исходя из шкалы релевантности документов: для результата агрегирования порог безразличия $\delta_c = 0{,}25$. Кроме того, были соблюдены ограничения, накладываемые на остальные значения порогов безразличия (в работе [16] показано, что значения порогов могут быть заданы так, что задача идентификации нечеткой меры не будет иметь решения, и предложены ограничения в виде неравенств, выполнение которых исключает подобную ситуацию).

Экспериментальные исследования проводились на статистически значимой выборке из 500 поисковых запросов, содержащих термины из обучающих примеров в различных комбинациях. В результате было установлено, что точность ранжирования результатов поиска при агрегировании на основе интеграла Шоке второго порядка улучшилась в среднем на 1 % по сравнению с точностью ранжирования при агрегировании средневзвешенным оператором. Здесь под точностью ранжирования понимается разность назначенной экспертом релевантности документа и релевантности, полученной на основе агрегирования зонных показателей с помощью одного из двух операторов.

Заключение. В статье рассмотрен вопрос практического применения нечетких мер и интеграла Шоке в области информационного поиска. Результаты экспериментов показали, что применяя в качестве оператора агрегирования зонных показателей интеграл Шоке по нечеткой мере, можно повысить точность ранжирования документов по сравнению с точностью ранжирования документов с помощью средневзвешенного оператора.

Предполагается исследовать использование предложенного метода определения весовых коэффициентов на различных коллекциях документов, а также практическую применимость рассматриваемого аппарата в других задачах информационного поиска (автоматическое исправление ошибок, автоматическое реферирование и аннотирование текстов).

СПИСОК ЛИТЕРАТУРЫ

1. Manning C., Raghavan P., Schutze H. Introduction to Information Retrieval // Cambridge University Press. 2008. — P. 544.
2. Robertson S., Zaragoza H., Taylor M. Simple BM25 Extension to Multiple Weighted Field // In ACM conference on Information Knowledge Management (CIKM) // http://www.hugo-zaragoza.net/academic/pdf/robertson_cikm04.pdf. 2004. P. 42—49. Дата обращения 13.05.2012.
3. Robertson S., Walker S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval // In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1994. — P. 232—241.
4. Cohen W., Singer Y. Context-sensitive Learning Methods for Text Categorization // ACM Transactions on Information Systems. 1999. 17(2). — P. 141—173.
5. Murata M., Ma Q., Uchimoto K., Ozaki H., Utiama M., Isahara H. Japanese Probabilistic Information Retrieval Using Location and Category Information // The Fifth International Workshop on Information Retrieval with Asian Languages. 2000. — P. 81—88.
6. Krysta S., Burges C. A Machine Learning Approach for Improved BM25 Retrieval // http://research.microsoft.com/pubs/101323/learningbm25msrtechreport.pdf. 2009. 25 p. (дата обращения 13.04.2012).
7. Marichal J.-L. An Axiomatic Approach to the Discrete Choquet Integral as a Tool to Aggregate Interacting Criteria // IEEE Transactions on Fuzzy Systems. 2000. № 8(6). — P. 800—807.
8. Grabisch M., Orlovski S., Yager R. Fuzzy Aggregation of Numerical Preferences// Handbook of Fuzzy Sets Series. Vol. 4: Fuzzy Sets in Decision Analysis, Operations Research and Statistics. — Dordrecht: Kluwer Academic, 1998. — P. 31—68.
9. Sugeno M. Theory of Fuzzy Integrals and its Applications: Ph.D. Thesis. — Tokyo. 1974. — 237 p.
10. Choquet G. Theory of capacities // Annales de l'Institut Fourier. 1953. № 5. — P. 131—295.
11. Fallah Tehrani A., Cheng W., Hüllermeier E. Preference Learning Using the Choquet Integral: The Case of Multipartite Ranking // IEEE Transactions on Fuzzy Systems // http://www.mathematik.unimarburg.de/~eyke/publications/draft_tfs11_choquet.pdf. 2012. 28 p. Дата обращения 10.06.2012.
12. Grabisch M., Kojadinovic I., Meyer P. A Review of Methods for Capacity Identification in Choquet Integral Based Multi-attribute Utility Theory: Applications of the Kappalab R package. 2008. № 2. — P. 766—785.
13. Mayag B., Grabisch M., Labreuche Ch. A Representation of Preferences by the Choquet Integral with Respect to a 2-additive Capacity. Theory and Decision. 2011. Vol. 71. — P. 297—324.
14. Kojadinovic I. Minimum Variance Capacity Identification // European Journal of Operational Research. 2007. № 177 (1). — P. 498–514.
15. Алфимцев А.Н., Лычков И.И. Метод обнаружения объекта в видеопотоке в реальном времени // Вестник ТГТУ. 2011. Т. 17. № 1. — С. 44—55.
16. Сакулин С.А. К вопросу об идентификации параметров интеграла Шоке 2-го порядка // Вестник ИРГТУ. 2008. № 3(35). — С. 205—208.

Статья поступила в редакцию 4.07.2012