НЕЙРОСЕТЕВОЙ МЕТОД СЕМАНТИЧЕСКОГО ВЕРОЯТНОСТНОГО ВЫВОДА В ЗАДАЧЕ УЛУЧШЕНИЯ РЕЛЕВАНТНОСТИ РЕЗУЛЬТАТОВ ПОИСКОВОГО ЗАПРОСА

М. Н. Калимолдаев, А. А. Пак*, С. С. Нарынов*

Институт проблем информатики и управления МОН РК,
Республика Казахстан, 050010, Алма-Ата,
*ТОО Alem Research, Республика Казахстан, 050010, Алма-Ата

УДК 004.8.032.26

Алгоритмы информационного поиска нацелены на получении наиболее релевантной выдачи документов по текстовому запросу. В большинстве прикладных семантических информационных систем пользователь для подготовки выборки документов производит итеративное уточнение параметров поискового запроса с целью улучшения релевантности документов для дальнейшего семантического анализа. Формирование качественного запроса из-за омонимической неоднозначности, большого разнообразия контекстов, значительной синонимичности слов и фраз является нетривиальной задачей. В языках поисковых запросов реализована грамматика логики высказываний. В данной статье предложен алгоритм уточнения поискового запроса, его подход основан на индуктивно-логическом выводе с использованием ручной бинарной классификации результатов первичной выдачи.

Ключевые слова: информационный поиск, нейронные сети, индуктивная логика.

The algorithms of information retrieval are aimed at getting the most relative documents' issue by the text query. In the majority of applied semantical informational systems an user itteratively makes the correction of search query parameters in order to improve the documents' relevance for further semantical processing. The construction of qualitive query is not simple task due to homonymic ambiguity, the big contextual variety, the significant synonymy of words and phrases. In the languages of search queries the first-order logic is used. In the paper the algorithm of search query correction is proposed. The main idea of the algorithm is based on inductive logical deduction and manual binary labeling of initial results of documents' issue.

Key words: natural language processing, neural networks, logical deduction.

Введение. Область исследования информационного поиска (ИП) широко освещена в русскоязычной и зарубежной литературе, о чем свидетельствуют многочисленные монографии и статьи. Термин „информационный поиск“ был предложен Кельвином Муэрсом в 1948 г. в его докторской диссертации, опубликован и употребляется в литературе с 1950 г. [1]. Центральным понятием ИП является характеристика релевантности — явно заданная величина, характеризующая пару документ-запрос, другими словами, насколько документ отвечает условиям запроса. Традиционно в моделях оценки релевантности используется статистический подход, который уже развивается более 30 лет. Наиболее значимыми трудами в данной области являются [2], [3], [4], [5], [6], [7], [8]. Первоначально область применения данных алгоритмов была ограничена библиотечным делом для наиболее быстрого и качественного поиска научной литературы. Широкое распространение ИП получил с появлением Интернета.

Важной частью ИП являются алгоритмы полнотекстового поиска. Существует целое семейство функций для формирования результирующей выборки по текстовому запросу. Во многих прикладных системах реализован алгоритм BM25 [9], [10]. Рассмотрим его подробнее. BM25 — это семейство поисковых функций на неупорядоченном множестве слов, так называемом „мешке слов“, и множестве документов. Алгоритм оценивает каждый документ из набора на основе частот встречаемости слов запроса, без учета близости между словами. Одна из распространенных форм этой функции описана ниже. Пусть дан запрос Q, содержащий слова $q_1 ... q_n$, тогда функция BM25 дает следующую оценку релевантности документа D запросу Q:

$$s(D,Q) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i,D)(k+1)}{f(q_i,D) + k(1 - b + b\frac{|D|}{<|D|>})}, \tag{1}$$

где $f(q_i, D)$ — частота слова q_i в документе D (следует отметить, что частоту слова можно также рассчитывать относительно группы документов, объеденных некоторым общим свойством), $|D|$ есть длина документа, $< |D| >$ — средняя длина документа, k и b — свободные коэффициенты.

Таким образом, оценка релевантности документа D подсчитывается на основе частот встречаемости каждого слова q_i, причем в топ выдачи попадают документы с более специализироваными, характерными словами. Информация, определенная в работе Шеннона, в контексте информационного поиска трудно измерима [11]. Оценки эффективности поисковой функции могут быть выполнены на основе двух характеристик, а именно точности и полноты, соответственно, точность определим следующим образом:

$$P = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|}, \tag{2}$$

где D_{rel} — множество релевантных документов в выдаче, а D_{retr} — множество документов, найденных системой. Следует отметить, что приведенная выше оценка является не единственной, к примеру, она не учитывает порядок выдачи или степень релевантности того или иного документа.

Вероятностный семантический вывод. Принцип семантического вероятностного вывода заключается в обнаружении максимально специфичных условных связей между n-граммами и документами. Информация, подаваемая на вход метода, кодируется одноместными предикатами $P_j^i(a) <=> (x_i(a) = x_i j)$, где $x_i(a)$ — информация, а $x_i j$ — ее значения на текущем объекте a, он представляет собой текст и метаинформацию документа. Вывод происходит замыканием относительно бинарных операторов $\neg, \wedge, \vee$ логики высказываний. Предикаты $P_j^i(a)$ и $\neg P_j^i(a)$ являются литералами (атомарными высказываниями или их отрицаниями), которые будем обозначать как $\alpha, \beta, \gamma, ... \in L$, где L — множество всех литералов в словаре $P_j^i, i = 1, ..., n; j = 1, ..., n_i$ [12]. В процессе семантического вероятностного вывода алгоритм обнаруживает множество R правил (условных связей) вида:

$$R = (\alpha_1 \wedge ... \wedge \alpha_k => \beta), \alpha_1, ..., \alpha_k, \beta \in L, \tag{3}$$

где $\alpha_1, ..., \alpha_k$ — входные предикаты, кодирующие вхождение n-грамм в текст. Правила характеризуются оценкой условной вероятности, которая вычисляется следующим образом.

Подсчитаем число случаев $n(\alpha_1, ..., \alpha_k, \beta)$, когда произошло событие $< \alpha_1, ..., \alpha_k, \beta >$ — одновременное срабатывание $< \alpha_1, ..., \alpha_k >$, иными словами, при $\beta, \neg\beta$. Далее подсчитаем случаи $n^+(\alpha_1, ..., \alpha_k, \beta)$ и $n^-(\alpha_1, ..., \alpha_k, \beta)$, раздельно при $\beta, \neg\beta$. Тогда оценка условной вероятности правила (1) равна:

$$\mu(\beta/\alpha_1, ..., \alpha_k) = \frac{(n^+(\alpha_1, ..., \alpha_k, \beta) - n^-(\alpha_1, ..., \alpha_k, \beta))}{(n(\alpha_1, ..., \alpha_k, \beta))}. \quad (4)$$

Алгоритм стремится максимизировать эту оценку. Правило $R_1 = (\alpha_1^1, ..., \alpha_{k1}^1 => \gamma)$ будем называть более общим, чем правило $R_2 = (\alpha_1^2, ..., \alpha_{k2}^2 => \gamma)$, обозначим это как $R_1 \succ R_2$ тогда и только тогда, когда $\{\alpha_1^1, ..., \alpha_{k1}^1\} \subset \{\delta_1^2, ..., \delta_{k2}^2\}, k1 < k2$, и не менее общим $R_1 \succeq R_2$, если $k1 \leq k2$. Очевидно, что $R_1 \succeq R_2 => R_1 \vdash R_2$ и $R_1 \succ R_2 => R_1 \vdash R_2$, где $\vdash$ — доказуемость в исчислении высказываний. Таким образом, не менее общие (и более общие) высказывания логически сильнее. Кроме того, более общие правила проще, так как содержат меньшее число литералов в посылке правила [13].

Вероятностным законом будем называть такое правило R, которое нельзя логически усилить, не уменьшив его условную вероятность, т.е. если $R' \succ R$, то $\mu(R') < \mu(R)$. Вероятностные законы — это наиболее общие, простые и логически сильные правила среди правил, имеющих не более высокую условную вероятность. Обозначим множество всех вероятностных законов через PL. Отношение вероятностного вывода $R_1 \sqsubseteq R_2, R_1, R_2 \in PL$ определим как одновременное выполнение двух неравенств $R_1 \succeq R_2$ и $\mu(R_1) < \mu(R_2)$. Если оба неравенства строгие, то отношение вероятносного вывода будем называть строгим отношением вероятностного вывода $R_1 \sqsubset R_2 <=> R_1 \succ R_2$ и $\mu(R_1) < \mu(R_2)$.

Семантическим вероятностным выводом будем называть максимальную (которую нельзя продолжить) последовательность вероятностных законов, находящихся в отношении строгого вероятностного вывода $R_1 \sqsubset R_2 \sqsubset ... \sqsubset R_k$. Последний вероятностный закон R_k в этом выводе будет называться максимально специфичным. Показано, что классификация по максимально специфическим правилам непротиворечива [13].

Алгоритм уточнения запроса. Уточнение запроса выполняется на основании ручной классификации первичной выдачи документов пользователем. Естественно предположить, что человек выполнит построение уточняющего запроса более качественно после изучения выборки документов. В данном случае алгоритм дает преимущество во времени построения запроса, основываясь на том предположении, что психофизическое восприятие подсветки ключевых слов в документе позволит быстрее определить релевантность документа, нежели совершить два действия, а именно определить релевантность и выделить стоп-слова. На рис. 1 представлена блок-схема алгоритма уточнения. Точкой входа в алгоритм является выдача документов первичного запроса. Вторым шагом алгоритма является классификация документов пользователем, а именно указываются те документы, наподобие которых нужно исключить из выдачи. На основе этого шага генерируется предикат β. Далее проводятся нормализация и взвешивание слов следующим образом, производятся стемминг и фильтрация стоп-слов, к которым относятся союзы, предлоги, междометия. Далее подсчитывается частота внутри класса документов, для которых выполняется β, обозначим $f(\beta)$ и $f(\neg\beta)$. Из набора термов исключаются все слова, для которых $\frac{(f(\beta))}{(f(\neg\beta))} < 3$. На основании полученной выборки формируется L, которое может быть дополнено литералами метаинформации документа, к примеру, url-ом или датой публикации. После чего к полученному множеству литералов и документов применяется

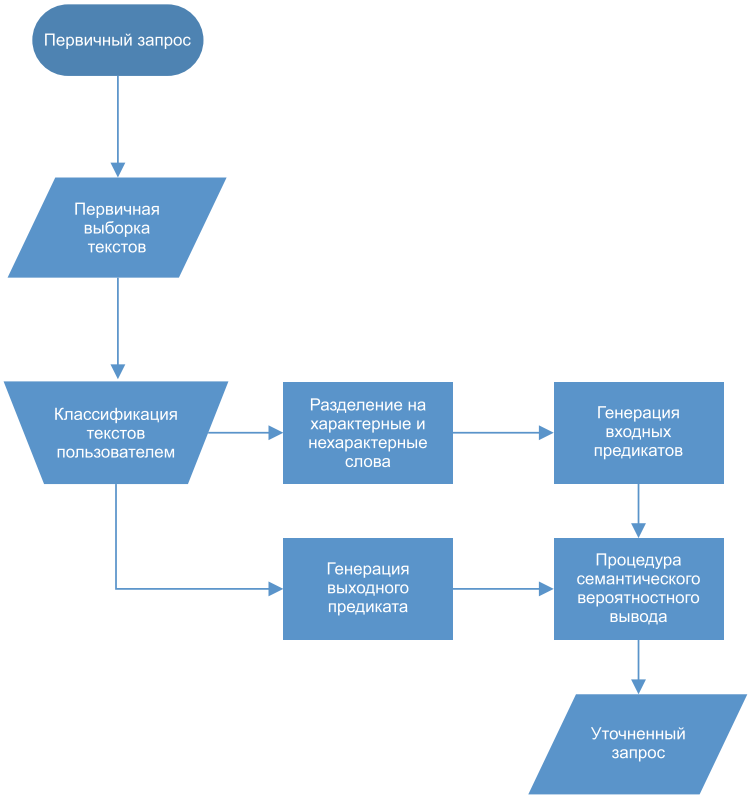

Рис. 1. Блок-схема алгоритма улучшения релевантности поиска

описанная выше процедура семантического вероятностного вывода. На выходе алгоритм возвращает несколько рекомендуемых наборов стоп-слов. Алгоритм был протестирован на выборке текстов, собранных из Интернета, объем составляет 2317 текстов. В качестве первичного запроса использовалось слово „замок“. Целью было выделить тексты, посвященные теме „замкóв“, и отфильтровать тексты про „зáмки“. Результирующий запрос выглядит следующим образом: „замок — (новые король расположен Рейтинг)“. Уточненный запрос повысил точность с P с 51,6 % до 91,6 %. Фраза „(новые король расположен Рейтинг)“ означает стоп-слова, т. е. в выдаче должны быть документы, в которых обязательно отсутствуют указанные слова.

Выводы. Таким образом, описанная выше модель является алгоритмом обучения с учителем, выявляет наиболее вероятные закономерности в данных в удобной для человека форме, а именно на языке логики высказываний. В применении к задаче уточнения поискового запроса метод показал быструю сходимость и практическую значимость. Метод внедрен и проходит апробацию в информационно-аналитической системе AlemSemantics.

Список литературы

1. Mooers C. The theory of digital handling of non-numerical information and its implications to machine economics // Proc. of the meeting of the Assoc. for Comp. Machinery at Rutgers University. 1950. New Jersey.

2. Maron M. E., Kuhns J. L. On relevance, probabilistic indexing and information retrieval // Journ. of the ACM. 1960. V. 7, N. 3. P. 216–244.

3. Robertson S. E. and Sparck Jones K. Relevance weighting of search terms // Journ. of the American Soc. for Information Science. 1977. V. 27. N. 3. P. 129–146.

4. Robertson S. E. and Walker S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval // Proc. of the 17th Annual Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval. 1994. P. 232–241.

5. Robertson S. E., Zaragoza H. and Taylor M. Simple BM25 extension to multiple weighted fields // Proc. of the 2004 ACM CIKM Intern. Conf. on Inf. and Knowledge Management. 2004. P 42–49.

6.Jones K.S., Walker S., Robertson S.E. A probabilistic model of information retrieval: development and comparative experiments // Inf. Process. Manage. N. 36(6). 2000. P. 779–808.

7. Jones K.S., Walker S., Robertson S.E. A probabilistic model of information retrieval: development and comparative experiments. Part 2 // Inf. Process. Manage. N. 36(6). 2000. P. 809–840.

8. Rijsbergen C.J. Information Retrieval. Second Edition. London: Butterworths, 1979.

9. Robertson S.E., Zaragoza H. The probabilistic relevance framework: BM25 and beyond //Foundation and Trends in information retrieval. 2009. V. 3. N. 4. P. 333–389.

10. Class BM25Similarity [электронный ресурс]. Режим доступа: http://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/BM25Similarity.html. (Дата обращения: 29.08.2014).

11. Shannon C.E., Weaver W. The Mathematical Theory of Communication. Urbana: University of Illinois Press, 1964.

12. E. E. Vityaev. Knowledge extraction from data. Computer Knowledge Model of cognitive process. Novosibirsk, 2006. P. 293.

13. Демин А. В., Витяев Е. Е. Логическая модель адаптивной системы управления // Нейроинформатика (электрон. журн.). 2008. Т. 3, № 1. С. 79–107.

Калимолдаев Максат Нурадилович — д-р физ.-мат. наук, проф., ген. дир., зав. лаб. математического моделирования и кибернетики Института проблем информатики и управления МОН РК, тел.: +7 (727) 272-37-11, e-mail: mnk@ipic.kz

Пак Александр Александрович — канд. техн. наук, ст. преп., рук. отд. науч. разраб. ТОО Alem Research, Республика Казахстан, тел.: +7 (701) 752-92-85, e-mail: aa.pak83@gmail.com

Нарынов Сергазы Сакенович — канд. техн. наук, ст. преп., ген. дир. ТОО Alem Research, Республика Казахстан, тел.: +7 (701) 723-01-62, e-mail: sergazy@gmail.com

Дата поступления — 01.09.2014