

Оценка эффективности применения контекстно-ассоциативных моделей текстов в задаче поиска по образцу на РОМИП'2005

© Беляев Д.В.

ЗАО «ОВИОНТ ИНФОРМ»
belyaev@oviont.ru

Аннотация

Статья посвящена применению контекстно-ассоциативных моделей анализируемых текстовых документов и основанного на них алгоритма выделения ключевых терминов в задаче уточнения поисковых запросов с обратной связью по релевантности с пользователями полнотекстовых поисковых систем. Экспериментальная проверка эффективности применения исследуемого алгоритма проводилась посредством участия в дорожке поиска по документу-образцу (mixed-feedback).

1. Введение

Несмотря на достаточно высокий современный уровень исследований, в теории информационного поиска остается ряд открытых проблем, одной из которых является проблема неоднозначного выбора терминов, вводимых пользователями информационно-поисковых систем (ИПС) в поисковых запросах. Эта проблема состоит в том, что пользователи ИПС часто применяют для описания ключевых понятий термины, отличные от терминов, которые используют авторы для описания тех же понятий в текстах.

Статистика показывает, что в общем случае два различных человека используют один и тот же термин для описания одного и того

же понятия менее чем в 20% случаев. Эта проблема стоит еще более остро для коротких поисковых запросов потому, что чем длиннее запрос, тем больше вероятность того, что наиболее важные термины из искомых документов попадут в поисковый запрос. Тем не менее, пользователи ИПС обычно не склонны выражать свою информационную потребность в виде длинных поисковых запросов или использовать специальные средства, позволяющие более точно формулировать запрос (например, языки поисковых запросов).

В большинстве случаев решение этой проблемы сводится к работе, связанной с итерационным переформулированием исходного запроса, которую пользователь ИПС берет на себя. Как следствие, методики уточнения поисковых запросов "вручную" в настоящее время играют ключевую роль для эффективного применения ИПС.

Основными недостатками многих известных методов уточнения поисковых запросов являются: жесткая привязка к моделям информационного поиска, реализованным в ИПС (например, в LSI и PhraseFinder); зависимость от качества исходного запроса (все методы, основанные на локальном контекстном анализе); необходимость постоянной актуализации тезаурусов или иных моделей коллекции документов, в которой осуществляется поиск (все методы, основанные на глобальном анализе).

Это обуславливает актуальность разработки и внедрения специальных методов, основанных на модели естественно-языковых текстов, которая бы позволила проводить их совместный анализ с целью выделения и оценки значимости общих ключевых терминов, что, в частности, позволило бы применять ее в алгоритмах уточнения поисковых запросов с обратной связью по релевантности

Целью настоящей работы является оценка эффективности применения метода уточнения поисковых запросов с использованием обратной связи по релевантности. Исследуемый метод независим от применяемой ИПС (в классе полнотекстовых ИПС, реализующих поиск по ключевым словам) и основан на контекстно-ассоциативных моделях текстовых документов, позволяющих выделять и оценивать значимость смысловых контекстов, определять ключевые термины и дать количественную оценку их смысловой значимости в анализируемом текстовом документе.

2. Контекстно-ассоциативная модель текста

Основная идея построения модели смысловых контекстов произвольного текстового документа состоит в выявлении устойчивых смысловых связей между терминами в различных предложениях

документа. Подробное описание предложенной модели смысловых контекстов, включающее доказательства ее основных свойств, приведено в работе [1].

Кратко суть предложенной модели можно представить следующим образом.

Произвольный текстовый документ d представляется в виде множества предложений Π^d , и множества терминов T^d , полученного в процессе терминологического разбора каждого из предложений текста. Под терминологическим разбором понимается выделение во входной последовательности слов предложения цепочек, образующих именные группы русского языка.

Отношения вхождения терминов в предложения описывается матрицей вхождения $\Delta^d = \|\delta_{ij}\|$.

Под множеством смысловых контекстов документа d понимается множество решений $C^d = \{\langle T, \Pi \rangle : T \subseteq T^d, \Pi \subseteq \Pi^d\}$ системы уравнений:

$$\begin{cases} \Pi = \text{Supp}(T) \\ T = \text{Cont}(\Pi) \end{cases}, \quad (1)$$

где операторы носителя терминов Supp и контента предложений Cont определяются следующим образом:

$$\text{Supp}(T) = \begin{cases} \bigcup_{t \in T} \Pi_t, & \text{если } T \neq \emptyset \\ \Pi^d, & \text{если } T = \emptyset \end{cases}$$

где $\Pi_t = \{\pi \in \Pi^d : \delta_{\text{ind}(t)\text{ind}(\pi)} = 1\}$ – носитель термина $t \in T^d$,

$$\text{Cont}(T) = \begin{cases} \bigcap_{\pi \in \Pi} T_\pi, & \text{если } \Pi \neq \emptyset \\ T^d, & \text{если } \Pi = \emptyset \end{cases}$$

где $T_\pi = \{t \in T^d : \delta_{\text{ind}(t)\text{ind}(\pi)} = 1\}$ – контент предложения $\pi \in \Pi^d$.

Пара $\langle T, \Pi \rangle$, являющаяся смысловым контекстом (т.е. удовлетворяющая уравнению (1)), обозначается $\llbracket T, \Pi \rrbracket$.

В работе [1] доказаны свойства смысловых контекстов, позволяющие определить понятия *образующих* смысловых контекстов и через них ввести операции объединения и пересечения смысловых контекстов.

Так, произвольное подмножество предложений $\Pi \subseteq \Pi^d$ или произвольное подмножество терминов $T \subseteq T^d$ однозначно задают некоторый соответствующий им смысловой контекст. Учитывая это, вводятся следующие обозначения смысловых контекстов:

$$\begin{aligned}\llbracket T \rrbracket &= \llbracket \text{ContSupp}(T), \text{Supp}(T) \rrbracket, \\ \llbracket \Pi \rrbracket &= \llbracket \text{Cont}(\Pi), \text{SuppCont}(\Pi) \rrbracket,\end{aligned}$$

где множества терминов T и предложений Π называются образующими соответствующих смысловых контекстов.

На множестве смысловых контекстов можно ввести отношение включения

$$\llbracket \Pi_1, T_1 \rrbracket \subseteq \llbracket \Pi_2, T_2 \rrbracket \Leftrightarrow \Pi_1 \subseteq \Pi_2,$$

а также операции объединения и пересечения смысловых контекстов (через объединения и пересечение их образующих):

$$\begin{aligned}\llbracket T_1, \Pi_1 \rrbracket \vee \llbracket T_2, \Pi_2 \rrbracket &= \llbracket \Pi_1 \cup \Pi_2 \rrbracket, \\ \llbracket T_1, \Pi_1 \rrbracket \wedge \llbracket T_2, \Pi_2 \rrbracket &= \llbracket \Pi_1 \cap \Pi_2 \rrbracket.\end{aligned}$$

Показано, что множество смысловых контекстов C^d документа d является замыканием множества базовых смысловых контекстов $C_1^d = \{ \llbracket \{\pi\} \rrbracket : \pi \in \Pi^d \} \cup \llbracket \emptyset \rrbracket$ относительно операции объединения смысловых контекстов $\vee$.

Для проведения количественной оценки значимости ассоциативных связей между смысловыми контекстами предложена контекстно-ассоциативная модель, являющаяся развитием ассоциативной модели, рассмотренной в работе [3].

Пусть $c = \llbracket T, \Pi \rrbracket \in C^d$. Первое предложение $\tilde{\pi} \in \Pi$ является порождающим предложением смыслового контекста c . Оставшаяся часть носителя смыслового контекста $\tilde{\Pi} = \Pi / \{\tilde{\pi}\}$ является областью существования смыслового контекста c .

Два смысловых контекста $c_\alpha, c_\beta \in C^d$ связаны в документе d непосредственной ассоциативной связью $\leftrightarrow$, если выполняется следующее условие:

$$c_\alpha \leftrightarrow c_\beta \Leftrightarrow \tilde{\Pi}_\alpha \cap \tilde{\Pi}_\beta \neq \emptyset.$$

В случае, когда контексты $c_\alpha, c_\beta \in C^d$ не связаны непосредственной ассоциативной связью, но имеется последовательность $c_j \in C^d$, $j = 1, 2, \dots, k$:

$$c_\alpha \leftrightarrow c_{j_1} \leftrightarrow c_{j_2} \leftrightarrow \dots \leftrightarrow c_{j_k} \leftrightarrow c_\beta,$$

то уровнем ассоциативной связи k между контекстами c_α и c_β называется наименьшая длина такой последовательности. Таким образом, непосредственная ассоциативная связь двух смысловых контекстов – это ассоциативная связь уровня 0.

Вес ассоциативной связи рассчитывается через ее уровень k :

$$w(c_\alpha, c_\beta) = 1 / 2^k.$$

Ассоциативной мощностью уровня l смыслового контекста $c \in C^d$ называется средневзвешенная сумма весов ассоциативных связей этого смыслового контекста с другими смысловыми контекстами документа d :

$$W^l(c) = \sum_{k=0}^l \left(\frac{1}{|C_c^k|} \sum_{c^* \in C_c^k} w(c, c^*) \right),$$

где C_c^k – множество контекстов, связанных с контекстом c ассоциативной связью уровня k .

Уровень l является эвристическим параметром контекстно-ассоциативной модели. Он может устанавливаться пользователем ИПС или выбираться экспериментальным путем.

3. Алгоритм уточнения поисковых запросов

Пусть D – конечное множество текстовых документов, Q – множество возможных поисковых запросов (в общем случае – бесконечное).

Пусть $q \in Q$ – исходный запрос, заданный пользователем ИПС, а $\tilde{D}_q$ – отклик ИПС на запрос q – упорядоченная последовательность документов из D (с точностью до перестановок документов с равными оценками релевантности).

Алгоритм уточнения поискового запроса:

1. Для всех документов $d \in \tilde{D}_q^{\text{rel}}$ строятся их контекстно-ассоциативные модели C^d .
2. Для всех терминов из анализируемых релевантных документов вычисляются весовые коэффициенты:

$$W(t, d) = \frac{1}{|C_t^d|} \sum_{c \in C_t^d} W^l(c),$$

где $C_t^d = \{[T, \Pi] \in C^d : t \in T\}$ — множество смысловых контекстов документа d , содержащих термин t .

3. Для всех терминов вычисляются обобщенные весовые коэффициенты терминов:

$$W(t) = \prod_{d \in \tilde{D}_q^{rel}} W(t, d).$$

4. Уточненный запрос строится в виде элементарной конъюнкции первых m терминов, имеющих наибольшие значения весовых коэффициентов. Параметр m является эвристическим параметром алгоритма уточнения запросов. Он может устанавливаться пользователем ИПС или выбираться экспериментальным путем.

Проводившиеся ранее экспериментальные оценки предложенного алгоритма с новостной коллекцией "Lenta.Ru" (23 805 txt-документов, 36 Мб) и коллекцией технической документации "CITForum.Ru" (26 911 html-документов, 455 Мб) осуществлялись при длине релевантной выборки от 1 до 6 документов, при этом уточненный запрос строился на основе различного числа ключевых терминов, значение которого также изменялось от 1 до 6, при фиксированном уровне ассоциативной сети, равном 1. При этом наилучшие результаты были получены при объеме релевантной выборки, составляющей 3 или 4 документа.

На основе этого был сделан вывод, что предложенная методика уточнения поисковых запросов не применима непосредственно к задаче поиска по документу-образцу, принадлежащему коллекции электронных документов. При этом лишь в случае 3 или 4 ключевых терминов достигается увеличение точности поиска на небольших значениях полноты (до 0,1-0,2). Это означает, что только среди первых результатов поиска наблюдается незначительное увеличение числа релевантных документов.

Однако, учитывая тот факт, что при поиске по документу-образцу заранее известен один заведомо релевантный документ (документ-образец), предложенный алгоритм можно сделать более эффективным, добавив в него "контроль за поведением документа-образца". Такой подход позволяет выбирать оптимальный поисковый запрос непосредственно в процессе поиска, варьируя число ключевых терминов, на основе которых строится уточненный запрос, и/или уровень ассоциативной сети.

Суть предлагаемого подхода состоит в том, что если положение документа-образца после поиска по уточненному запросу ухудшилось (т.е. он оказался дальше от начала результата поиска, по сравнению с результатом поиска по исходному запросу), то можно попытаться подобрать оптимальные значения числа ключевых терминов и/или уровня ассоциативной сети, при которых документ-образец оказывается как можно ближе к началу отклика ИПС. При этом уточненный запрос соответствует наилучшему положению документа-образца в отклике ИПС, в противном случае исходный запрос считается неуточняемым при помощи предложенного метода.

4. Постановка экспериментов

Целью экспериментальной проверки предложенного метода уточнения поисковых запросов является оценка эффективности его применения к решению задачи уточнения поисковых запросов методом обратной связи по релевантности с пользователем ИПС на основе 1 релевантного документа на тестовой коллекции электронных документов, тематический состав и объем которой достаточно близок к соответствующим параметрам реальных коллекций.

Задачи экспериментальной оценки включают:

1. Оценку эффективности непосредственного применения предложенного метода уточнения поисковых запросов (без "контроля за поведением документа-образца").
2. Проверку предположения о возможности увеличения эффективности метода при помощи "контроля за поведением документ-образцом".
3. Оценку оптимальных параметров работы метода (числа ключевых терминов – Term, на основе которых строится уточненный запрос, и уровня контекстно-ассоциативной сети – Level).
4. Сравнительную оценку эффективности работы метода на коллекциях электронных документов различной тематики (при поиске в web, при поиске в коллекции нормативно-правовых документов и при поиске в смешанной коллекции).

Исходным набором тестовых данных являлось объединение коллекции html-документов "Narod.Ru" и коллекции нормативно-правовых документов «Legal» (предоставленная компанией «Кодекс»).

Набор тестовых заданий включал 8205 пар тестовых запросов вида "исходный запрос и один релевантный документ". Задания строились на основе набора запросов, который оценивался в

РОМИП'2003 и РОМИП'2004. Релевантным документом для каждого запроса считался документ, оцененный, как строго релевантный, то есть все ассесоры признали его таковым [4].

Для проведения экспериментов в качестве тестовой ИПС использовалась shareware-версия Yandex.Server Standard 3.2.4 (<http://company.yandex.ru/technology/products/yandex-server.xml>).

Электронные документы тестовой коллекции были представлены для индексирования в виде файлового архива, при этом из него были предварительно исключены файлы, не содержащие в себе ни одного слова на русском языке в кодировке Win 1251 (что допустимо по условиям РОМИП).

В ходе экспериментальной проверки было осуществлено 3 эксперимента по уточнению запросов:

- без контроля за поведением документа-образца при фиксированном значении уровня контекстно-ассоциативной сети;
- с контролем за поведением документа-образца при фиксированном значении уровня контекстно-ассоциативной сети;
- с различными значениями уровня контекстно-ассоциативной сети при фиксированном значении числа ключевых терминов в уточненном запросе.

Параметры, при которых проводились прогоны, приведены в табл. 1.

Табл. 1

№ эксп.	Контроль за документом-образцом	Число ключевых терминов в запросе	Уровень контекстно-ассоциативной сети	Индексы прогонов
1	-	1-6	1	01-06
2	+	1-6	1	11-16
3	+	3	0-4	20-24

Помимо указанных прогонов, с целью проведения сравнительной оценки, в качестве базового прогона был выполнен поиск по исходному запросу (прогон с индексом 00).

Кроме официальных метрик РОМИП, для оценки качества отклика ИПС по запросу q использовалась интегральная оценка отклика:

$$quality(\tilde{D}_q, \tilde{D}_q^{rel}) = \sum_{d \in \tilde{D}_q^{rel}} \frac{1}{ind(d)},$$

где $\tilde{D}_q^{rel}$ – множество документов, релевантных запросу q , $ind(d)$ – позиция документа в отклике ИПС. Также оценивалось изменение интегральной оценки откликов по всему множеству тестовых запросов Q , выраженное в процентах от ее исходного значения:

$$\Delta quality = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{quality(\tilde{D}_{q_i}^*, \tilde{D}_{q_i}^{rel}) - quality(\tilde{D}_{q_i}, \tilde{D}_{q_i}^{rel})}{quality(\tilde{D}_{q_i}, \tilde{D}_{q_i}^{rel})}, \quad (2)$$

где q^* – уточненный запрос.

При этом для любого отдельно взятого запроса значение $\Delta quality > 0$ означало, что запрос был уточнен, в противном случае запрос уточнен не был.

5. Оценка результатов экспериментов

Ниже приводятся результаты предварительного анализа проведенных экспериментов.

Как и ожидалось, непосредственное применение алгоритма уточнения поисковых запросов приводит к ухудшению результатов поиска при любых значениях параметров Term и Level. Результаты этого эксперимента (эксперимент №1) проиллюстрированы на рис. 1 и приведены в табл. 2а, 3а и 4а (для различных текстовых коллекций). Подобный результат можно объяснить тем, что предложенный метод уточнения поисковых запросов предполагает совместный анализ нескольких релевантных документов, полученных в ходе поиска по исходному запросу. Предварительные эксперименты показывают, что оптимальная длина релевантной выборки составляет, как правило, 3 или 4 документа.

Гипотеза о том, что использование информации о положении документа-образца при поиске по исходному и уточненному запросу позволит осуществить "настройку" параметров работы алгоритма экспериментально подтверждается. Из результатов эксперимента №2, проиллюстрированных на рис. 2 и приведенных табл. 2а, 3а и 4а видно, что уже в случае, когда уточненный запрос строится лишь на одном ключевом термине документа-образца, достигается увеличение средней оценки качества поиска по критерию изменения интегральной оценки качества поиска.

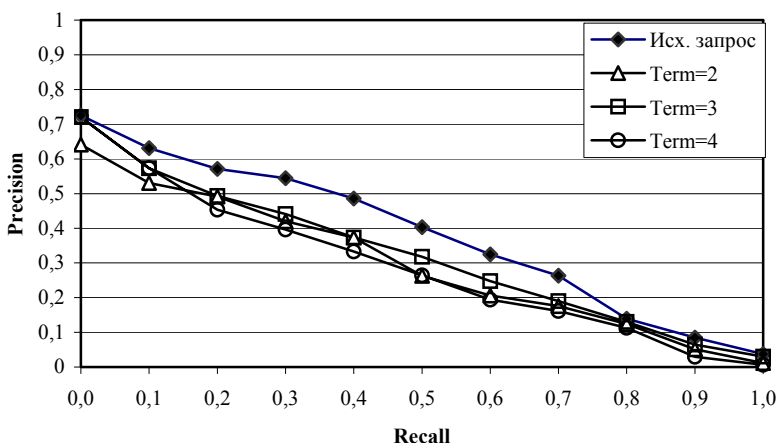


Рис. 1. Эксперимент без контроля за документом-образцом
(Narod.Ru + Legal, OR / Relevant-Minus, Term=2..4)

При этом оценки оптимального значения числа ключевых терминов, сделанные на основе критерия "точность/полнота" поиска и на основе изменения интегральной оценки практически совпадают, а наилучшая эффективность работы метода уточнения поисковых запросов достигается при 2-3 ключевых терминах (2 в случае оценки на основе критерия "точность/полнота" и 3 – на основе интегральной оценки). Таким образом, исходя из того, что положение документа-образца непосредственно связано с интегральной оценкой, можно сделать вывод о правомерности применения этого критерия для оценки изменения качества поиска.

Сравнительный анализ применимости предлагаемого метода к коллекциям различных типов (рис. 3) показал, что при незначительном отличии в проценте уточненных запросов существенное улучшение качества поиска достигается при поиске в Web (более чем на 25% – для коллекции Narod.Ru и примерно на 8% – для коллекции Legal).

Это можно объяснить тем, что предметная область, к которой относятся нормативно-правовые документы, является существенно более формализованной и обладает хорошо устоявшейся терминологией, что влечет хорошее качество исходных запросов. (Этот результат согласуется с предварительными экспериментами с новостной коллекцией и коллекцией, содержащей техническую документацию.) По той же причине для эффективного уточнения запросов

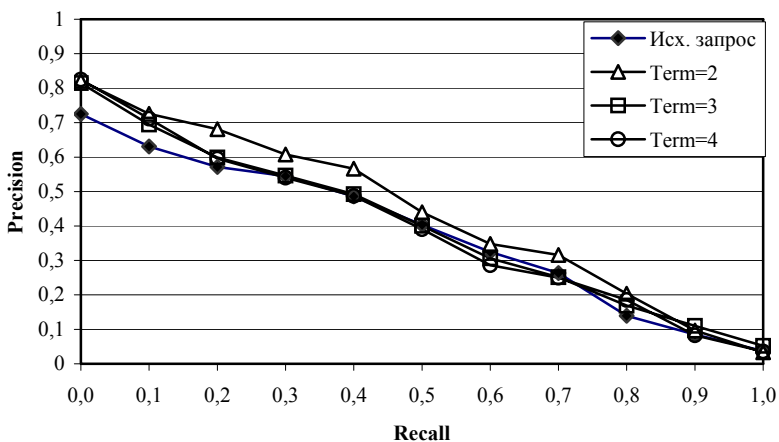


Рис. 2. Эксперимент с контролем за документом-образцом
(Narod.Ru + Legal, OR / Relevant-Minus, Term=2..4)

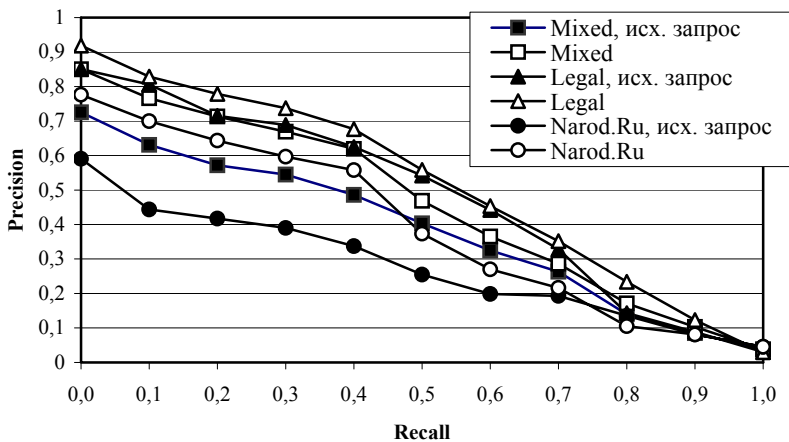


Рис. 3. Сравнительная оценка уточнения запросов для различных типов
текстовых коллекций (OR / Relevant-Minus)

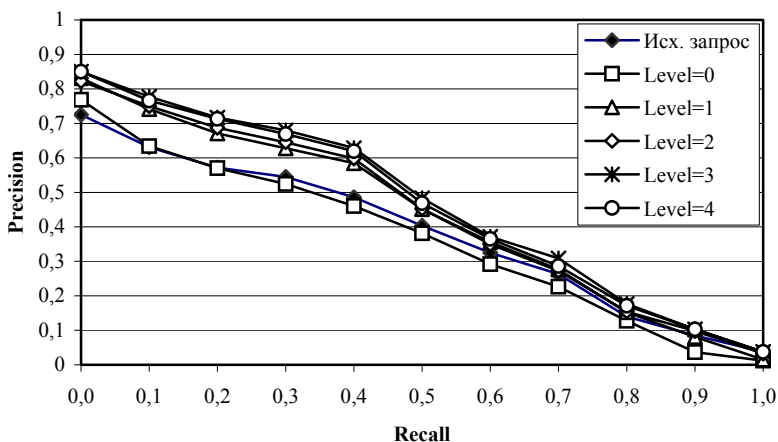


Рис. 4. Эксперимент с различными значениями уровня контекстно-ассоциативной сети (Narod.Ru + Legal, OR / Relevant-Minus)

при поиске в Web требуется использовать не один, а несколько документов образцов, что позволит более точно выделять ключевые термины.

Оценка влияния уровня контекстно-ассоциативной сети на результат применения предложенного метода уточнения запросов с контролем положения документа-образца показан на рис. 4. Соответствующие результаты, основанные на интегральной оценке качества поиска, приведены в табл. 2б, 3б и 4б.

Видно, что добавление в контекстно-ассоциативную сеть ассоциативных связей более высокого уровня приводит к увеличению эффективности работы алгоритма уточнения поисковых запросов. При этом наблюдается быстрое замедление роста качества поиска, начиная с уровня Level=2. С учетом того, что вычислительная сложность алгоритма увеличивается с ростом уровня контекстно-ассоциативной сети, можно предположить, что не имеет смысла применять контекстно-ассоциативные сети с ассоциативными связями выше 2 уровня.

Табл. 2, а

Тип эксперимента	Без контроля по документу-образцу						С контролем по документу-образцу					
Число ключевых терминов	1	2	3	4	5	6	1	2	3	4	5	6
Общее число запросов	58											
Тип коллекции	Narod.Ru + Legal / and (relevant-minus)											
Число улучшенных запросов	12	22	24	23	20	19	17	23	24	24	22	20
Процент улучшенных запросов	20,7%	37,9%	41,4%	39,7%	34,5%	32,8%	29,3%	39,7%	41,4%	41,4%	37,9%	34,5%
Начальная средняя оценка качества	1,9287											
Конечная средняя оценка качества	0,9797	1,3923	1,7070	1,7243	1,6982	1,6203	2,0073	2,0650	2,0499	2,0266	1,9851	1,9508
Средний процент изменения качества	-49,2%	-27,8%	-11,5%	-10,6%	-12,0%	-16,0%	4,1%	7,1%	6,3%	5,1%	2,9%	1,1%
Тип коллекции	mixed / or (relevant-minus)											
Число улучшенных запросов	14	25	24	24	23	22	18	25	25	25	24	22
Процент улучшенных запросов	24,1%	43,1%	41,4%	41,4%	39,7%	37,9%	31,0%	43,1%	43,1%	43,1%	41,4%	37,9%
Начальная средняя оценка качества	2,2873											
Конечная средняя оценка качества	1,2718	1,8284	2,0829	2,0598	1,9733	1,9041	2,4360	2,5074	2,4823	2,4544	2,4176	2,3705
Средний процент изменения качества	-44,4%	-20,1%	-8,9%	-9,9%	-13,7%	-16,8%	6,5%	9,6%	8,5%	7,3%	5,7%	3,6%

Табл. 2, б

Уровень ассоциативной сети	0	1	2	3	4	0	1	2	3	4
Тип коллекции	Narod.Ru + Legal / and (relevant-minus)					Narod.Ru + Legal / or (relevant-minus)				
Общее число запросов	58									
Число улучшенных запросов	17	24	26	26	26	19	25	26	26	26
Процент улучшенных запросов	29,3%	41,4%	44,8%	44,8%	44,8%	32,8%	43,1%	44,8%	44,8%	44,8%
Начальная средняя оценка качества	1,9287					2,2873				
Конечная средняя оценка качества	2,0199	2,0499	2,0768	2,1184	2,1184	2,3407	2,4823	2,5296	2,5296	2,5210
Средний процент изменения качества	4.7%	6.3%	7.7%	9.8%	9.8%	2.3%	8.5%	10.6%	10.6%	10.2%

Табл. 3, а

Тип эксперимента	Без контроля по документу-образцу					С контролем по документу-образцу						
Число ключевых терминов	1	2	3	4	5	6	1	2	3	4	5	6
Общее число запросов	28											
Тип коллекции	Narod.Ru / and (relevant-minus)											
Число улучшенных запросов	6	12	14	12	10	9	8	13	14	13	12	10
Процент улучшенных запросов	21,4%	42,9%	50,0%	42,9%	35,7%	32,1%	28,6%	46,4%	50,0%	46,4%	42,9%	35,7%
Начальная средняя оценка качества	0,8611											
Конечная средняя оценка качества	0,4132	0,6272	0,7457	0,6966	0,6497	0,5666	0,9096	1,0107	1,0438	1,0413	1,0269	0,9029
Средний процент изменения качества	-52,0%	-27,2%	-13,4%	-19,1%	-24,5%	-34,2%	5,6%	17,4%	21,2%	20,9%	19,3%	4,9%
Тип коллекции	web / or (relevant-minus)											
Число улучшенных запросов	8	15	14	13	13	12	9	15	15	14	14	12
Процент улучшенных запросов	28,6%	53,6%	50,0%	46,4%	46,4%	42,9%	32,1%	53,6%	53,6%	50,0%	50,0%	42,9%
Начальная средняя оценка качества	1,6039											
Конечная средняя оценка качества	1,0182	1,5304	1,5243	1,3916	1,2197	1,1545	1,9219	2,0174	2,0091	2,0134	1,9226	1,7183
Средний процент изменения качества	-36,5%	-4,6%	-5,0%	-13,2%	-24,0%	-28,0%	19,8%	25,8%	25,3%	25,5%	19,9%	7,1%

Табл. 3, б

Уровень ассоциативной сети	0	1	2	3	4	0	1	2	3	4
Тип коллекции	Narod.Ru / and (relevant-minus)					Narod.Ru / or (relevant-minus)				
Общее число запросов	28									
Число улучшенных запросов	9	14	15	15	15	11	15	15	15	15
Процент улучшенных запросов	32,1%	50,0%	53,6%	53,6%	53,6%	39,3%	53,6%	53,6%	53,6%	53,6%
Начальная средняя оценка качества	0,8611					1,6039				
Конечная средняя оценка качества	0,9507	1,0438	1,0457	1,0457	1,0457	1,8707	2,0091	2,0496	2,0496	2,0496
Средний процент изменения качества	10,4%	21,2%	21,4%	21,4%	21,4%	16,6%	25,3%	27,8%	27,8%	27,8%

Табл. 4, а

Тип эксперимента	Без контроля по документу-образцу						С контролем по документу-образцу					
Число ключевых терминов	1	2	3	4	5	6	1	2	3	4	5	6
Общее число запросов	30											
Тип коллекции	Legal											
Число улучшенных запросов	6	10	10	11	10	10	9	10	10	11	10	10
Процент улучшенных запросов	20,0%	33,3%	33,3%	36,7%	33,3%	33,3%	30,0%	33,3%	33,3%	36,7%	33,3%	33,3%
Начальная средняя оценка качества	2,9251											
Конечная средняя оценка качества	1,5085	2,1064	2,6043	2,6834	2,6767	2,6037	3,0319	3,1718	3,1102	2,9960	2,9792	2,9648
Средний процент изменения качества	-48,4%	-28,0%	-11,0%	-8,3%	-8,5%	-11,0%	3,6%	8,4%	6,3%	2,4%	1,8%	1,4%

Табл. 4, б

Уровень ассоциативной сети	0	1	2	3	4
Тип коллекции	Legal				
Общее число запросов	30				
Число улучшенных запросов	8	10	11	11	11
Процент улучшенных запросов	26,7%	33,3%	36,7%	36,7%	36,7%
Начальная средняя оценка качества	2,9251				
Конечная средняя оценка качества	3,0232	3,1102	3,1718	3,1718	3,1718
Средний процент изменения качества	3,4%	6,3%	8,4%	8,4%	8,4%

6. Выводы

Экспериментальная оценка показала, что предложенный алгоритм уточнения поисковых запросов может применяться в задаче поиска по документу-образцу, принадлежащему коллекции электронных документов. При этом для его эффективного применения необходим контроль за положением документа-образца при поиске по исходному и уточненному запросу.

Проведена сравнительная оценка применимости предложенного метода для уточнения запросов при поиске в коллекции нормативно-правовых документов и в Web. Ее результаты показали относительно хорошую эффективность применения предложенного метода при поиске в текстовых коллекциях общей тематической направленности.

В ходе экспериментов проанализировано влияние параметров предложенной контекстно-ассоциативной модели на эффективность работы алгоритма уточнения поисковых запросов и даны рекомендации по выбору:

- уровня контекстно-ассоциативной сети (не более 2);
- числа ключевых терминов, используемых при построении уточненных запросов (от 2 до 4).

Литература

- [1] Беляев Д.В. Ассоциативная модель смысловых контекстов и ее применение в задаче уточнения поисковых запросов.// Электронный журнал "Труды МАИ". - 2005, N18 – http://www.mai.ru/projects/mai_works/articles/num18/article9/auther.htm
- [2] Беляев Д.В. Контекстный анализ естественно-языковых текстов в задаче уточнения поисковых запросов. – В кн.: Материалы XIV Международной конференции по вычислительной механике и современным прикладным программным системам (ВМСППС-2005). – М.: Вузовская книга, 2005. – 482 с.
- [3] Чанышев О.Г. Ассоциативная модель естественного текста.// Вестник ОмГУ. – 1997, N4. – с. 17-20.
- [4] ROMIP Web site, 2005. <http://romip.narod.ru>

Estimation of application efficiency of contextually-associative texts models in pattern-matching problem on ROMIP'2005

Dmitry V. Belyaev

This article presents an application of contextually-associative model of texts and the algorithm of key terms allocation in analyzed text documents in a problem of relevance-feedback query specification. Within ROMIP'2005 experimental check of application efficiency of investigated algorithm was estimated on pattern-matching problem (mixed-feedback).