Контекстно-ассоциативный метод уточнения поисковых запросов и аннотирования текстовых документов

© Дмитрий Беляев

ОВИОНТ ИНФОРМ
belyaev@oviont.ru

Аннотация

В статье рассматривается метод решения задачи уточнения поисковых запросов с обратной связью по релевантности с пользователями информационно-поисковых систем (ИПС) и метод построения контекстно-зависимых аннотаций текстов на естественном языке. В основе обоих методов лежит контекстно-ассоциативный подход к анализу значимости терминов и фрагментов естественно-языковых текстов. Приводятся результаты участия в семинаре РОМИП'2006 в рамках дорожек поиска по документу-образцу и контекстно-зависимого аннотирования текстовых документов.

1. Введение

Ассоциативные модели текстовых документов представляют в настоящее время достаточно распространенный аппарат анализа текстов на естественном языке в силу относительной простоты их построения. Главным образом они применяются в задачах, решение которых не требует детального грамматического анализа текстов, а основным критерием информативности служит частота повторения терминов (слов и словосочетаний) в различных фрагментах текстовых документов [1,2].

Рассматриваемый подход, основанный на использовании контекстно-ассоциативных моделей текстовых документов [3,4], был исследован ранее в ходе участия в семинаре РОМИП в 2005 году [5].

Полученные результаты показали применимость контекстно-ассоциативных моделей для различных типов коллекций текстовых документов при решении задачи поиска по документу образцу – частному случаю более общей проблемы уточнения поисковых запросов с обратной связью по релевантности с пользователями ИПС. При этом были получены оценки значений параметров метода поиска по документу-образцу, которые позволили достичь результатов, приемлемых с точки зрения поставленной задачи.

Одним из существенных недостатков контекстно-ассоциативных моделей является значительная вычислительная сложность их построения и анализа, что сказывается на ограниченной применимости основанных на них методов анализа текстов большого объема. Таким образом, одной из целей участия в РОМИП'2006 стала практическая оценка подхода, который, с одной стороны, позволил бы уменьшить объем контекстно-ассоциативных моделей и тем самым увеличить скорость их анализа, а с другой – сохранить их свойства.

Решение задачи оценки значимости терминов в анализируемых документах в контексте исходного запроса лежит в основе метода уточнения поисковых запросов. Однако с использованием ассоциативных моделей текстовых документов можно решать и "двойственные" задачи, состоящие в выделении в текстах наиболее значимых фрагментов [2]. Оценка применимости контекстно-ассоциативных моделей текстов к задачам контекстно-зависимого аннотирования явилась второй целью участия в семинаре РОМИП'2006.

2. Контекстно-ассоциативные модели текстов

2.1 Общий подход

В отчете об участии в РОМИП'2005 и ряде других статей [3-5] приведено достаточно подробное описание контекстно-ассоциативных моделей текстов и рассмотрены их свойства.

Под смысловым контекстом документа d понимается пара $c = [\![T, \Pi]\!]$, состоящая из подмножества терминов документа T и подмножества предложений документа Π, удовлетворяющая системе уравнений:

$$\begin{cases} \Pi &= \mathrm{Supp}(T) \\ T &= \mathrm{Cont}(\Pi) \end{cases}, \qquad (1)$$

где оператор носителя $\mathrm{Supp}(T)$ и оператор контента $\mathrm{Cont}(\Pi)$ задают, соответственно, множество предложений, включающих термины T и множество терминов, входящих в предложения Π.

При этом множество смысловых контекстов C^d документа d строится через замыкание множества базовых смысловых контекстов

$$C_1^d \stackrel{\Delta}{=} \left\{ [\![\{\pi\}]\!] : \pi \in \Pi^d \right\} \cup [\![\varnothing]\!] \qquad (2)$$

относительно операции их объединения.

Считается, что два смысловых контекста $c_\alpha, c_\beta \in C^d$ связаны в документе d непосредственной ассоциативной связью $\leftrightarrow$, если выполняется условие

$$c_\alpha \leftrightarrow c_\beta \quad \Leftrightarrow \quad \widetilde{\Pi}_\alpha \cap \widetilde{\Pi}_\beta \neq \varnothing ,$$

а в случае, когда контексты $c_\alpha, c_\beta \in C^d$ не связаны непосредственной ассоциативной связью (т.е. ассоциативной связью уровня 0), но имеется последовательность $c_j \in C^d$, $j = 1,2,\ldots k$:

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

то уровнем ассоциативной связи k между контекстами c_α и c_β называется наименьшая длина такой последовательности.

Вес ассоциативной связи рассчитывается через ее уровень k:

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

и задает оценку значимости контекста c в анализируемом документе:

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

где l – уровень контекстно-ассоциативной сети.

2.2 Особенности реализации

Рассмотренный подход к построению контекстно-ассоциативной сети применим к произвольным текстовым документам. Однако если текстовые документы имеют большой объем, то получаемая модель является слишком громоздкой.

Для смешанной коллекции РОМИП (Narod.Ru + Legal) на множестве документов из тестовых заданий 2006 года были получены следующие оценки (табл. 1):

Табл. 1

Объем документа, $\left\|\Pi^d\right\|$	Среднее число смысловых контекстов, $\left\|C^d\right\|$	Число непосредственных ассоциативных связей
5	9	29
10	24	149
25	101	1371
50	365	7636

Как видно из табл. 1, линейное увеличение объема документа приводит экспоненциальному росту числа ассоциативных связей, что сказывается на увеличении вычислительных затрат.

2.3 Модификация контекстно-ассоциативной модели

Для снижения эффекта лавинообразного роста контекстно-ассоциативной модели с увеличением объема анализируемого документа предлагается подход, позволяющий рассматривать лишь те смысловые контексты, которые являются значимыми с точки зрения исходного запроса.

В рамках этого подхода рассматривается не все множество C^d решений уравнения 1, а его подмножество, которое строится на основе множества базовых смысловых контекстов, соответствующих терминам, входящим в исходный запрос:

$$C_1^q \overset{\Delta}{=} \left\{ [\![\{t\}]\!] : t \in T^q \cup \overline{T}^q \right\} \cup [\![\varnothing]\!], \qquad (3)$$

где T^q – множество терминов, входящих в запрос q, $\overline{T}^q$ – множество терминов, встречающихся с терминами из T^q в различных предложениях документа d.

Можно показать, что такое ограничение на множество базовых смысловых контекстов задает подмножество множества C^d, состоящее только из тех смысловых контекстов, которые включают термины из запроса или находятся с ними в непосредственной ассоциативной связи. При этом операция объединения смысловых контекстов вводится не через объединение множеств образующих их предложений, а через множества образующих их терминов.

Оценки размеров контекстно-ассоциативных моделей, полученных на множестве пар «запрос-документ» из тестовых заданий 2006 года приведены в табл. 2.

Табл. 2

Объем документа, $\left\|\Pi^d\right\|$	Среднее число смысловых контекстов, $\left\|C^d\right\|$	Число непосредственных ассоциативных связей
5	6	11
10	15	67
25	31	227
50	71	729

Видно, что предлагаемый подход к построению контекстно-ассоциативных моделей позволяет в среднем уменьшить число ассоциативных связей более чем в 10 раз для документов, содержащих более 50 предложений.

3. Поиск по документу-образцу

3.1 Подготовка и проведение экспериментов

Набор заданий включал 11358 пар запросов вида «запрос + релевантный документ». Задания строились на основе набора запросов, которые оценивались на семинарах РОМИП 2003-2005 гг.

Набором тестовых данных являлось объединение коллекции web-документов (Narod.ru) и коллекции нормативно-правовых документов (Legal). 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).

Для проведения экспериментов в качестве тестовой ИПС использовалась shareware-версия Яндекс.Server Standard 3.4.6 (http://company.yandex.ru/technology/products/yandex-server.xml).

Для поиска по документу-образцу применялся алгоритм, использовавшийся в 2005 году [3], но основанный на модифицированном подходе к построению контекстно-ассоциативных моделей. В ходе выполнения заданий было осуществлено 3 прогона:

- 1 прогон – «эталонный» поиск по исходному запросу;
- 2 прогон – поиск по уточненному запросу, полученному на основе исходного запроса и документа-образца с применением модифицированных контекстно-ассоциативных моделей;
- 3 прогон – поиск по уточненному запросу с контролем за положением документа-образца (относительно его положения в

результатах 1-го прогона) с применением модифицированных контекстно-ассоциативных моделей и автоматическим выбором наилучших значений параметров метода: число ключевых терминов, вводимых в запрос, варьировалось от 1 до 5, уровень контекстно-ассоциативной сети – от 0 до 2.

3.2 Результаты поиска по документу-образцу

На момент написания статьи были получены результаты оценки только для тестовой коллекции Narod.Ru, которые проводились на 233 заданиях вида «запрос + релевантный документ» и включавших 20 различных поисковых запросов. Другими участниками РОМИП'2006 результаты для заданий на основе запросов к коллекции Narod.Ru сданы не были.

Среди результатов, полученных от организаторов РОМИП, были рассмотрены таблицы релевантности, оцененные при «глубине» котла 50 (когда каждый из попавших в котел документов оценивался экспертами на соответствие запросу) при строгих (AND, рис. 1 и 2) и нестрогих (OR, рис. 3 и 4) требованиях к релевантности.

Сравнительные результаты изменения точности для различных уровней полноты поиска по исходному и уточненному запросам, полученным в ходе участия в РОМИП в 2005 и 2006 годах, приведены в табл. 3.

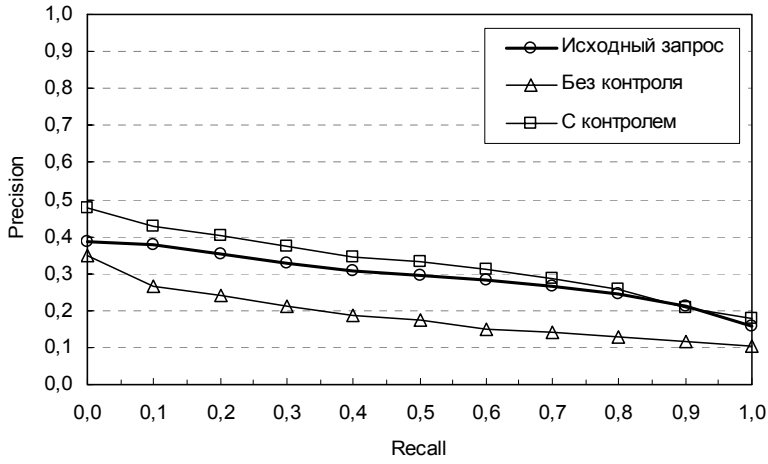

Рис. 1. РОМИП 2006, поиск по образцу в web (AND, pd50)

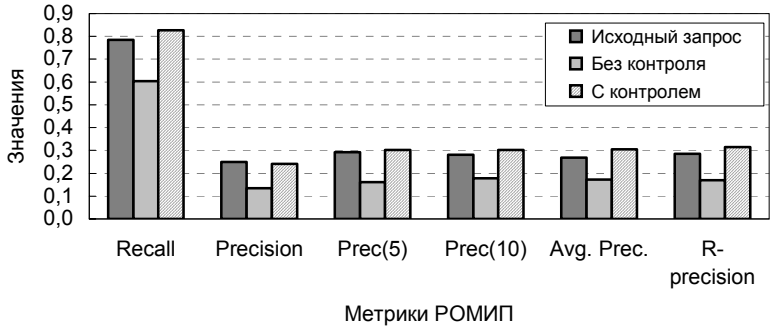

Рис. 2. Метрики РОМИП, поиск по образцу в web (AND, pd50)

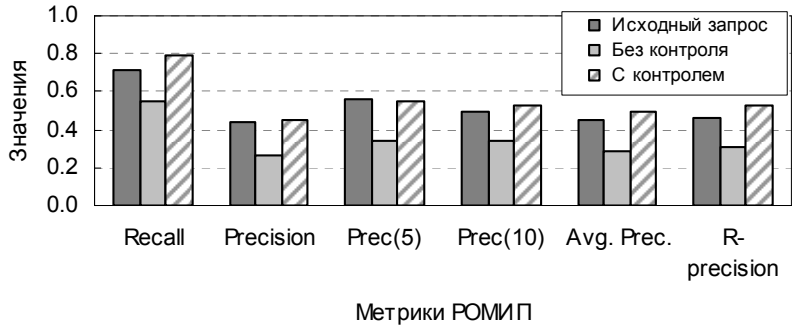

Рис. 3. Метрики РОМИП, поиск по образцу в web (OR, pd50)

Табл. 3

Recall	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
Narod.Ru, AND, pd50											
Δ Prec., 2006	24%	12%	15%	15%	13%	13%	10%	9%	5%	-1%	15%
Δ Prec., 2005	22%	40%	38%	34%	30%	44%	-2%	-11%	-4%	14%	90%
	2%	-28%	-23%	-19%	-17%	-31%	**12%**	**20%**	**9%**	-15%	-75%
Narod.Ru, OR, pd50											
Δ Prec., 2006	9%	8%	15%	11%	14%	15%	11%	7%	8%	12%	21%
Δ Prec., 2005	24%	49%	45%	40%	52%	18%	21%	-12%	-15%	-14%	-21%
	-15%	-41%	-30%	-29%	-38%	-3%	-10%	**19%**	**23%**	**26%**	**42%**

Из таблицы видно, что применение модифицированных контекстно-ассоциативных моделей в алгоритме уточнения поисковых запросов дает положительный результат, однако при этом наблюдается некоторое снижение эффективности его работы с точки зрения изменения качества поиска по исходному и уточненному запросам.

Таким образом, применение модифицированных контекстно-ассоциативных моделей позволяет получить заметный выигрыш в скорости работы алгоритма, без существенного снижения эффективности решения задачи уточнения поисковых запросов.

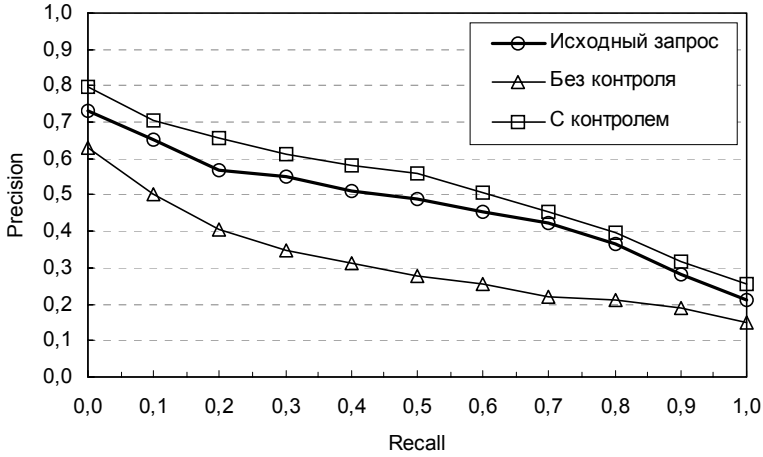

Рис. 4. РОМИП 2006, поиск по образцу в web (OR, pd50)

4. Контекстно-зависимое аннотирование

4.1 Алгоритм построения аннотаций

Для аннотирования текстовых документов на основе контекстно-ассоциативных моделей решалась задача, «двойственная» задаче оценки значимости терминов:

1. Аннотируемый документ d после удаления всех тэгов и замены escape-последовательностей на соответствующие символы (где это было возможно) рассматривается как plan text, разбитый на предложения.

2. Для аннотируемого документа d с учетом поискового запроса q на основе множества базовых смысловых контекстов, задаваемых выражением (3), строится его модифицированная контекстно-ассоциативная модель C^q.

3. Для всех предложений π, входящих в модель C^q, вычисляются весовые коэффициенты:

$$W(\pi, q) = \frac{1}{\left|C_\pi^q\right|} \sum_{c \in C_\pi^q} W^l(c),$$

где $C_\pi^q = \{[\![T, \Pi]\!] \in C^q : \pi \in \Pi\}$ – множество смысловых контекстов, содержащих предложение π.

Исключение составляли предложения, ограниченные тэгами <title> и <h1>, которым априори присваивались наибольшие весовые коэффициенты в случае, если они содержали термины из запроса.

4. В аннотацию включается *m* предложений, имеющих наибольшие значения весовых коэффициентов. Параметр *m* может выбираться исходя из ограничений на максимальный размер аннотации.

В силу того, что по условиям РОМИП'2006 при формировании аннотации ее максимальная длина не должна превышать 300 символов, на 4 шаге алгоритма в каждом предложении (начиная с самого весомого) выделялся фрагмент на основе подхода, описанного в статье [6]. При этом работа алгоритма начиналась с наиболее значимых предложений в порядке убывания их весовых коэффициентов, а при окончательной «сборке» аннотации предложения размещались в порядке их следования в реферируемом документе. Если в ходе работы алгоритма предложение обрезалось, то для удобства чтения аннотации в местах обрезки добавлялись троеточия: «...».

Критерием окончания работы алгоритма построения аннотации являлось либо достижение аннотации ограничения в 300 символов, либо отсутствие в очередном добавляемом предложении ключевых терминов.

4.2 Подготовка и проведение экспериментов

Выполнение задания, состоящего из 42587 пар вида «запрос + документ» проводилось на основе тестовых коллекций web-документов (Narod.Ru) и нормативно-правовых документов (Legal).

Для разбиения документов на предложения использовался пакет Text::Parser языка Perl.

Ниже приводятся примеры аннотаций документов, вошедших в оцениваемую выборку (правописание сохранено):

Запрос (sum2625): «отправка сообщений на мтс» Документ: http://boykoa.narod.ru/mobile.htm (1,2 Кб)
...сдесь можно отправить SMS сообщения клиентам компании МТС. Сотовая связь. СОТОВАЯ СВЯЗЬ... http://www.mts.ru/sms/ - сдесь можно отправить SMS сообщения клиентам компании МТС. Они подключены по коду +7 902. Так... Так же сдесь находится остальная информация компании МТС, их реклама, предложения...
Ключевые слова: сообщение, сотовый, МТС, компания, телефон, код, клиент, абонент

Запрос (sum8240): «горный университет» Документ: http://zahdima.narod.ru/VUZ/VUZ.HTM (682 Кб)
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ГОРНЫЙ УНИВЕРСИТЕТ (МГИ) 117935... Высшие, средние Высшие, средние специальные и начальные учебные заведения Москвы и Московской области Указатели специальностей. Программы...
Ключевые слова: университет, горный, обучение, Москва, Московская область, учебный, заведение, средний, программа

Запрос (sum39160): «расчет пособий по временной нетрудоспособности в 2003 году» Документ: 901852036 (46,7 Кб)
...Об утверждении Порядка назначения и... Об утверждении Порядка назначения и осуществления страховых выплат по обязательному социальному страхованию от несчастных случаев на производстве и... Нетрудоспособными считаются: лица, не достигшие 18 лет, а также старше 18 лет, обучающиеся в учебных...
Ключевые слова: обязательное социальное страхование, пособие, профессиональный заболевание, несчастный случай, выплата

4.3 Результаты контекстно-зависимого аннотирования

При оценке релевантности аннотаций использовались метрики РОМИП [7], имеющие следующие обозначения:

AAccur – доля релевантных аннотаций (по мнению экспертов), которым соответствуют реально релевантные документы;

AError – доля нерелевантных аннотаций, которым соответствуют релевантные документы;

PrecA – доля аннотаций признанных релевантными.

PrecisionD – доля документов, для которых строились аннотации, которые были признаны релевантными при оценке в 2004 году (по таблице релевантности для adhoc дорожек).

Результаты были рассчитаны для четырех комбинаций требований к релевантности аннотаций/документов: and/and, and/or, or/and, or/or. При этом результаты работы экспертов объединялись на основе различных уровней порога согласованности: Vital, Relevant-plus и Relevant-minus [7].

Результаты оценки аннотирования документов приведены в табл. 4. Через «/» обозначены наилучшие значения, полученные по результатам прогонов всех участников (всего 2 участника):

Табл. 4

Метрика	AAccur	AError	PrecA (macro)	PrecA	PrecD (macro)	PrecD
Vital						
and/and	**.958 / .958**	.505 / .503	.096 / .103	.131 / .148	.565 / .566	**.557 / .557**
and/or	**.983 / .983**	.635 / .631	.096 / .103	.131 / .148	**.676 / .676**	.668 / .669
or/and	**.840 / .840**	**.427 / .427**	.273 / .319	.279 / .325	.565 / .566	**.557 / .557**
or/or	**.939 / .939**	.553 / .545	.273 / .319	.279 / .325	**.676 / .676**	.668 / .669
Relevant-plus						
and/and	**.878 / .878**	.466 / .463	.180 / .205	.189 / .218	.565 / .566	**.557 / .557**
and/or	**.956 / .956**	.598 / .594	.180 / .205	.189 / .218	**.676 / .676**	.668 / .669
or/and	**.802 / .802**	.363 / .356	.411 / .473	.401 / .469	.565 / .566	**.557 / .557**
or/or	**.915 / .915**	.471 / .456	.411 / .473	.401 / .469	**.676 / .676**	.668 / .669
Relevant-minus						
and/and	**.848 / .848**	.414 / .409	.288 / .321	.292 / .327	.565 / .566	**.557 / .557**
and/or	**.949 / .949**	**.531 / .531**	.288 / .321	.292 / .327	**.676 / .676**	.668 / .669
or/and	**.768 / .768**	**.299 / .299**	.519 / .603	.510 / .598	.565 / .566	**.557 / .557**
or/or	**.894 / .894**	**.374 / .374**	.519 / .603	.510 / .598	**.676 / .676**	.668 / .669

Как видно из таблицы, результаты, полученные различными участниками, отличаются несущественно. При этом представленный метод дает хорошие результаты по критерию AAccur. Это означает, что большинству релевантных (по мнению экспертов) аннотаций соответствуют реально релевантные документы.

По числу нерелевантных аннотаций, которым соответствуют релевантные документы (критерий AError) наблюдается отставание от лучших результатов, за исключением оценок, полученных при использовании порога vital.

4. Заключение

В представленной работе рассматривалась модифицированная контекстно-ассоциативная модель естественно-языковых текстовых документов и оценивалась ее применимость к решению задач уточнения поисковых запросов (в рамках дорожки поиска по документу-образцу) и контекстно-зависимого аннотирования.

Была достигнута основная цель, состоящая в снижении вычислительных затрат, связанных с использованием контекстно-ассоциативных сетей при анализе текстовых документов. Как показали эксперименты, применение модифицированных контекстно-ассоциативных моделей текстов позволяет получить существенный выигрыш в производительности без заметного снижения качества поиска по уточненным запросам.

Литература

[1] *Ермаков А.Е., Плешко В.В*. Ассоциативная модель смысла текста в прикладных задачах компьютерного анализа полнотекстовых документов. // Русский язык: исторические судьбы и современность. Международный конгресс. Труды и материалы. – Москва: МГУ, 2001.

[2] *Чанышев О.Г.* Ассоциативная модель естественного текста. // Вестник ОмГУ, N4, 1997. – с. 17-20.

[3] *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.

[4] *Беляев Д.В.* Экспериментальная проверка применения контекстно-ассоциативных моделей в задаче уточнения поисковых запросов. // Информационные технологии и программирование: Межвузовский сборник статей. Вып. 2 (14) – М.: МГИУ, 2005. – с. 19--30.

[5] *Беляев Д.В.* Оценка эффективности применения контекстно-ассоциативных моделей текстов в задаче поиска по образцу на РОМИП'2005. – В кн.: Труды третьего российского семинара РОМИП'2005 (Ярославль, 6 октября 2005 г.) – Санкт-Петербург: НИИ Химии СПбГУ, 2005. – с. 89-105.
[6] *Браславский П., Колычев И.* eXtragon: экспериментальная система для автоматического реферирования веб-документов. – В кн.: Труды третьего российского семинара РОМИП'2005 (Ярославль, 6 октября 2005 г.) – Санкт-Петербург: НИИ Химии СПбГУ, 2005. – с. 40-53.
[7] Результаты для дорожки контекстно-зависимого аннотирования. Приложение G. – Труды третьего российского семинара РОМИП'2005 (Ярославль, 6 октября 2005 г.) – Санкт-Петербург: НИИ Химии СПбГУ, 2005. – с. 222-223.

Context-associative method for query specification and texts summarization on ROMIP'2006

© Dmitry Belyaev
belyaev@oviont.ru

This article introduces the method of solving a query specification problem with relevance-feedback to the users of IR-systems and method of text summarization for texts in natural language. Both methods are based on context-associative approach to the analysis of importance of terms and fragments of textual documents. The results of participation in ROMIP'2006 in pattern-based adhoc track and query-based summarization track are given.