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

Беляев Дмитрий Владимирович
Каф. «Математическая кибернетика», МАИ
ЗАО «ОВИОНТ ИНФОРМ»
Belyaev@oviont.ru

Актуальность задачи уточнения запросов

- Пользователи ИПС и авторы искомых документов используют **различные термины** для описания одних и тех же понятий

  до 80% случаев.

- Пользователи ИПС используют **короткие запросы**

  в более 50% случаев – 2-4 слова (1-2 термина).

- Пользователи ИПС **неточно выражают цель поиска** (т.е. свою информационную потребность) в виде поискового запроса в силу

  различий между языком запросов ИПС и ЯПП предметной области.

- **Пользователи вынуждены видоизменять запрос для достижения более высоких результатов поиска.**

  это происходит в более чем 90% случаев.

Документ $d = \langle T^d, \Pi^d, \mathrm{Inc} \rangle \in D,$

где Inc – отношение вхождения терминов в предложения.

Носитель (Supp) терминов $T \subseteq T^d$ – множество предложений, включающих эти термины.

$$\mathrm{Supp}(T) = \begin{cases} \bigcap_{t \in T} \Pi_t, & \text{если } T \neq \varnothing, \\ \Pi^d, & \text{если } T = \varnothing, \end{cases} \qquad \Pi_t = \left\{ \pi \in \Pi^d : \delta_{\mathrm{ind}(t)\,\mathrm{ind}(\pi)} = 1 \right\}$$

Контент (Cont) предложений $\Pi \subseteq \Pi^d$ – множество терминов, входящих в каждое из этих предложений.

$$\mathrm{Cont}(\Pi) = \begin{cases} \bigcap_{\pi \in \Pi} T_\pi, & \text{если } \Pi \neq \varnothing, \\ T^d, & \text{если } \Pi = \varnothing, \end{cases} \qquad T_\pi = \left\{ t \in T^d : \delta_{\mathrm{ind}(t)\,\mathrm{ind}(\pi)} = 1 \right\}$$

Смысловые контексты

Множество смысловых контекстов документа d :

$$C^d = \left\{ [[T, \Pi]] : T \subseteq T^d, \Pi \subseteq \Pi^d \right\},$$

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

В ассоциативных моделях текстов

$$\langle t, \Pi \rangle,$$

где $\Pi = \mathrm{Supp}(t)$

- вхождения терминов в предложения текста

$\rightarrow$

В контекстно-ассоциативной модели текстов

$$\langle T, \Pi \rangle,$$

- смысловой контекст: устойчивые сочетания терминов и предложений текста

Контекстно-ассоциативная модель текста

$c = [[T, \Pi]] \in C^d$ $\quad \tilde{\pi} \in \Pi$ - порождающее предложение

$\tilde{\Pi} = \Pi / \{\tilde{\pi}\}$ - область существования

Непосредственная ассоциативная связь контекстов:

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

Ассоциативная связь уровня k:

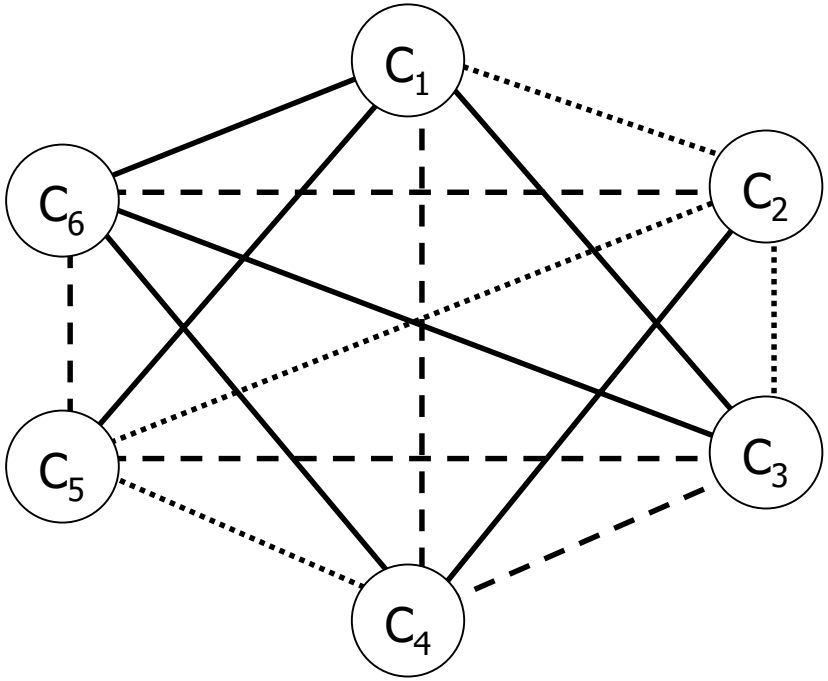
$$c_\alpha \leftrightarrow c_{j_1} \leftrightarrow c_{j_2} \leftrightarrow \ \ldots \ \leftrightarrow c_{j_k} \leftrightarrow c_\beta$$

Ассоциативные связи:

0-го уровня ——————

1-го уровня – – – – – – –

2-го уровня .....



Оценка значимости смысловых контекстов

Вес ассоциативной связи:

$w(c_\alpha, c_\beta) = 1/2^k$, где k – уровень ассоциативной связи

Ассоциативная мощность смысловых контекстов уровня k :

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

где C_c^i - множество контекстов, связанных с контекстом c ассоциативной связью уровня i.

Алгоритм построения уточненного запроса

$\widetilde{D}_q^{\text{rel}}$

Построение контекстно-ассоциативных моделей для $d \in \widetilde{D}_q^{\text{rel}}$

Расчет весовых коэффициентов

$$W(t,d) = \frac{1}{n(t,d)} \sum_{c_j \in C_t^d} W^k(c_j),$$

$$C_t^d = \{c \in C^d : t \in [T] \equiv c\}, \quad n(t,d) = \left|C_t^d\right|.$$

Расчет обобщенных весовых коэффициентов

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

$$t_1 \quad t_2 \quad \dots \quad t_m \quad | \quad t_{m+1} \quad \dots$$

$$W(t_1) \geq W(t_2) \geq \dots \geq W(t_m) \geq W(t_{m+1}) \geq \dots$$

$t_1 \,\&\, t_2 \,\&\, \dots \,\&\, t_m$

Проведение экспериментальной оценки

Условия проведения экспериментов:

- дорожка поиска по документу-образцу
- коллекция Mixed = Nerod.Ru + Legal
- тестовые задания вида: запрос + 1 релевантный документ
- поисковая система - Yandex.Server Standard 3.2.4

Критерии оценки:

- официальные метрики РОМИП
- интегральная оценка отклика по положению релевантных документов:

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

- изменение интегральной оценки отклика по всему множеству запросов:

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

Задачи экспериментальной оценки

1. Оценка эффективности непосредственного применения предложенного метода уточнения поисковых запросов.

2. Проверка предположения о возможности увеличения эффективности метода при помощи "контроля за поведением документа-образца".

3. Выбор оптимальных параметров работы метода:
   - числа ключевых терминов в уточненном запросе;
   - уровня контекстно-ассоциативной сети.

4. Проверка правомерности использования критерия $\Delta quality$ для оценки результативности процедуры уточнения запросов.

5. Сравнительная оценка эффективности работы метода на коллекциях электронных документов различной тематики.

Параметры экспериментов

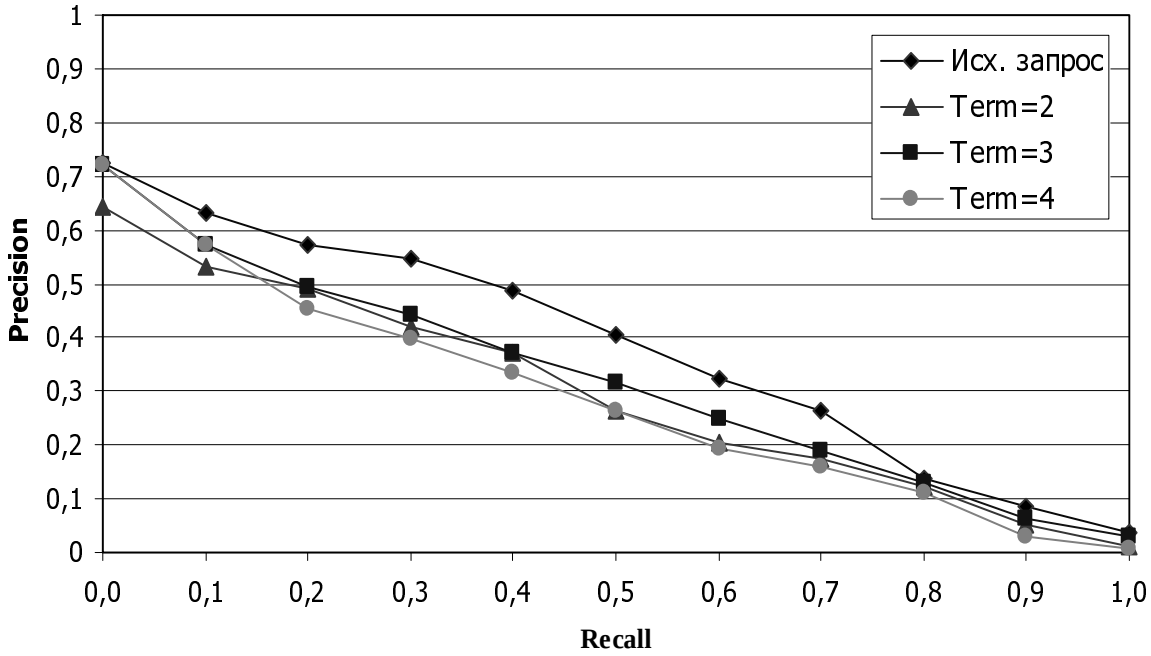
Проведенные эксперименты:

- 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.

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

Непосредственное применение метода

Одного документа-образца недостаточно для непосредственного применения предложенного метода уточнения запросов.



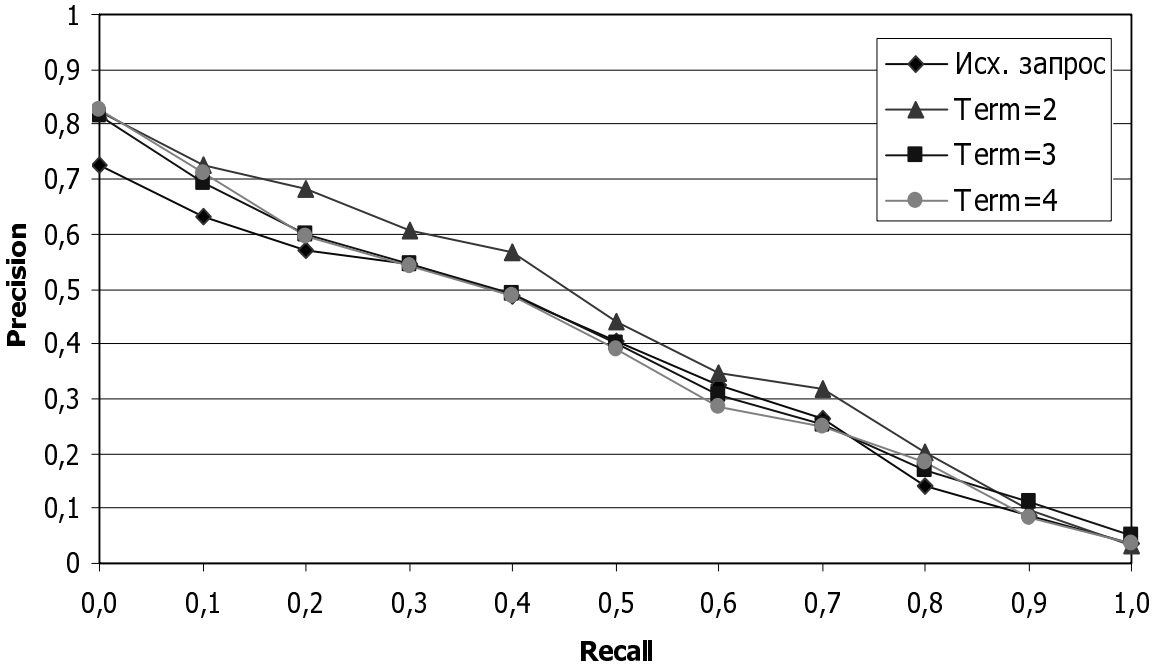
Число ключевых терминов:	**1**	**2**	**3**	**4**	**5**	**6**
Общее число запросов:	58					
Число улучшенных запросов:	14	25	24	24	23	22
Процент улучшенных запросов:	24,1%	43,1%	41,4%	41,4%	39,7%	37,9%
Оценка $quality$ по исходному запросу:	2,287					
Оценка $quality$ по уточненным запросам:	1,271	1,828	2,082	2,059	1,973	1,904
Оценка изменения $\Delta quality$	-44,4%	-20,1%	-8,9%	-9,9%	-13,7%	-16,8%

Эксперимент с подбором числа ключевых терминов

Выбор оптимального числа ключевых терминов в запросе позволяет добиться улучшения качества поиска.

Рекомендуемое число ключевых терминов: от 2 до 4

Narod.Ru + Legal, OR / Relevant-Minus



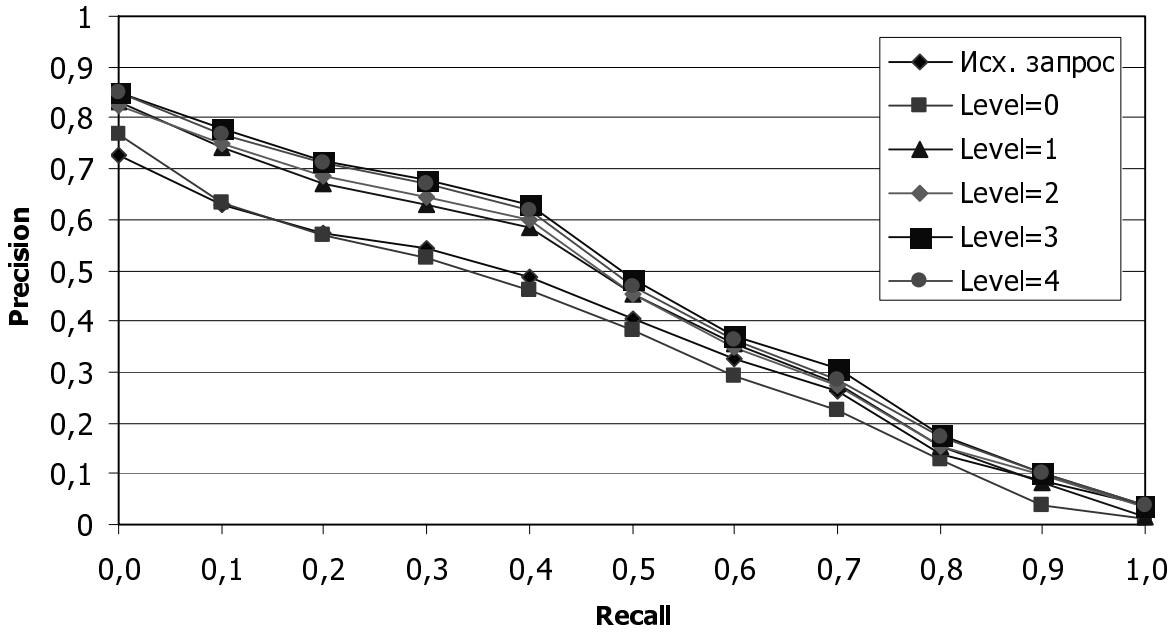
Число ключевых терминов:	**1**	**2**	**3**	**4**	**5**	**6**
Общее число запросов:	58					
Число улучшенных запросов:	18	25	25	25	24	22
Процент улучшенных запросов:	31,0%	43,1%	43,1%	43,1%	41,4%	37,9%
Оценка $quality$ по исходному запросу:	2,287					
Оценка $quality$ по уточненным запросам:	2,436	2,507	2,482	2,454	2,417	2,370
Оценка изменения $\Delta quality$	6,5%	9,6%	8,5%	7,3%	5,7%	3,6%

Эксперимент с уровнем ассоциативной сети

Добавление ассоциативный связей приводит к увеличению качества поиска.

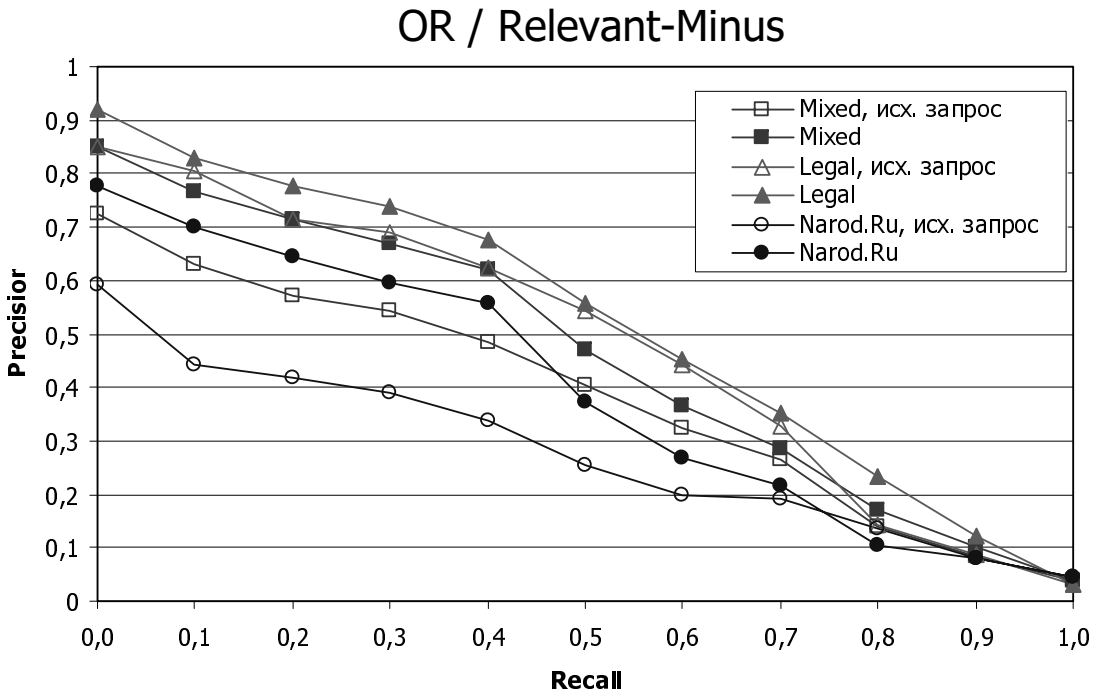
Рекомендуемый уровень ассоциативной сети: 2 или 3

Narod.Ru + Legal, OR / Relevant-Minus



Число ключевых терминов:	**0**	**1**	**2**	**3**	**4**
Общее число запросов:	58				
Число улучшенных запросов:	19	25	26	26	26
Процент улучшенных запросов:	32,8%	43,1%	44,8%	44,8%	44,8%
Оценка $quality$ по исходному запросу:	2,287				
Оценка $quality$ по уточненным запросам:	2,340	2,482	2,529	2,529	2,521
Оценка изменения $\Delta quality$	2,3%	8,5%	10,6%	10,6%	10,2%

Оценки для различных типов коллекций



Коллекция:	**Midex**			**Legal**			**Narod.Ru**		
Запрос:	**исх.**	**опт.**	**%**	**исх.**	**опт.**	**%**	**исх.**	**опт.**	**%**
Recall	0,486	0,648	33,2%	0,572	0,704	23,1%	0,395	0,587	48,8%
Precision(5)	0,603	0,690	14,3%	0,760	0,793	4,4%	0,436	0,579	32,8%
Average precision	0,376	0,451	20,1%	0,474	0,529	11,5%	0,270	0,368	36,2%
Precision(10)	0,548	0,645	17,6%	0,700	0,740	5,7%	0,386	0,543	40,7%
R-precision	0,398	0,500	25,5%	0,495	0,556	12,4%	0,294	0,439	49,0%
Precision	0,416	0,378	-9,0%	0,467	0,377	-19,1%	0,361	0,379	5,1%

Выводы

- Предложенный метод может применяться в задаче поиска по документу-образцу, принадлежащему коллекции электронных документов, при условии "контроля за положением документа-образца" и эмпирическом подборе оптимальных значений параметров.

- Метод эффективно работает на тестовых коллекциях общей тематической направленности.

- Анализ влияния параметров контекстно-ассоциативной модели на качество работы метода позволил дать рекомендации по выбору:
  - уровня контекстно-ассоциативной сети (от 2 до 3);
  - числа ключевых терминов (от 2 до 4).