УДК 519.766.2

Ассоциативная модель смысловых контекстов и ее применение в задаче уточнения поисковых запросов

Д.В. Беляев

В статье дается обзор методов, применяемых для уточнения (расширения) информационно-поисковых запросов – одной из ключевых проблем теории информационного поиска. Вводится формализованное понятие смыслового контекста. Предлагается модель смысловых контекстов для текстов на естественном языке. Рассматриваются свойства предложенной модели, и обосновывается метод ее построения. Приводятся алгоритмы построения и применения ассоциативной модели смысловых контекстов для решения задачи уточнения информационно-поисковых запросов методом обратной связи по релевантности с пользователями информационно-поисковых систем.

Введение

Одной из основных проблем полнотекстового информационного поиска является проблема неоднозначного выбора терминов, используемых 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 [1]. 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 [2]. 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) [3].

В связи с этим одними из ключевых задач в теории информационного поиска стали:

- **задача расширения поисковых запросов**, состоящая в добавлении в исходный запрос пользователей ИПС синонимов или словоформ ключевых терминов запроса, без изменения смыслового содержания исходного запроса;
- **задача уточнения (или переформулирования) поисковых запросов**, состоящая в изменении исходного запроса посредством учета ключевых слов из релевантных (т.е. соответствующих информационным потребностям пользователей ИПС) документов с целью уточнения смыслового содержания запроса и, как следствие, точности поиска.

Методы решения этих задач можно разделить на два класса:

- **методы автоматического уточнения (расширения) запросов**, не требующие получения в ходе своей работы дополнительной информации от пользователей ИПС;
- **методы, использующие обратную связь с пользователем**, в ходе работы которых пользователь должен предоставить дополнительную информацию, позволяющую осуществить более точный поиск.

В зависимости от объема используемой информации, методы уточнения запросов подразделяются на **глобальные** и **локальные**.

Глобальные методы основаны на использовании информации обо всей коллекции документов, в которой осуществляется поиск. Одним из первых глобальных методов уточнения запросов является метод кластеризации терминов [4]. Другими наиболее известными глобальными методами уточнения запросов являются: методы, использующие тематическую кластеризацию документов [5], скрытое семантическое индексирование (LSI) [2, 6], технологии PhraseFinder и аналогичные ей [7, 8], основанные на построении тезаурусов посредством автоматического выделения терминов и устойчивых словосочетаний. Главное отличие глобальных методов состоит в том, что для их применения необходима предварительная обработка всей коллекции документов на этапе ее индексирования.

Идея использования для уточнения запросов документов, полученных в ходе поиска по исходному запросу, легла в основу локальных методов уточнения запросов [9], использующих принцип псевдообратной связи по релевантности [10, 11], а также методов, в которых применяются вероятностные подходы к определению ключевых терминов из релевантных документов [6, 12].

Однако, наилучшие результаты уточнения запросов дают методы, использующие обратную связь с пользователем ИПС [13]–[17], работающие в несколько итераций. Среди результатов поиска по исходному запросу пользователю ИПС требуется указать заинтересовавшие его документы, в ходе дальнейшего анализа которых и строится уточненный запрос.

Принципиально новые подходы вызваны появлением ИПС с большим числом пользователей, таких как Yahoo, Google, Рамблер, Яндекс, что позволяет накапливать и анализировать статистику по вводимым запросам и документам, которые пользователь выбирает для более детального изучения, выражая тем самым свою информационную потребность [18].

Выбор подхода к уточнению запросов зависит от многих факторов, например, числа релевантных запросу пользователя документов, объема коллекции документов, предрасположенности пользователя к поиску в несколько итераций (с использованием обратной связи) [6, 19].

В основе большинства методов, решающих задачи расширения и уточнения поисковых запросов, независимо от используемого подхода, лежат модели текстовых документов, характеристики которых оказывают непосредственное влияние на выбор ключевых терминов [17]. В настоящей статье предлагается подход, основанный на использовании формальной модели смысловых контекстов, позволяющий перейти от анализа слов и терминов, составляющих документ, к анализу смысловых контекстов – устойчивых сочетаний групп терминов, несущих в анализируемом документе единую смысловую нагрузку.

1. Формальное определение понятия смыслового контекста

Предлагаемая модель смысловых контекстов основана на идее выявления устойчивых смысловых связей между терминами в различных документах [16]. Однако, в отличие от рассмотренного подхода, анализ будет проводиться не во всей коллекции документов, а в каждом документе в отдельности.

Рассмотрим произвольный текстовый *документ* (или его некоторый сегмент – главу, раздел, подраздел и т.п.). Так как минимальным фрагментом текста с законченным смысловым содержанием в общем случае является предложение, рассмотрим анализируемый документ d как последовательность предложений $\langle \pi_1, \pi_2, \ldots, \pi_n \rangle$ и представим его в виде множества предложений

$$\Pi^d = \{\pi_1, \pi_2, \ldots, \pi_n\}, \quad (1)$$

где n – число предложений в документе d.

Под *терминами* будем в дальнейшем понимать слова или словосочетания, обозначающие в документе d некоторые сущности. Пусть

$$T^d = \{t_1, t_2, \ldots, t_m\}, \quad (2)$$

– *словарь документа* d объемом m терминов. В дальнейшем будем считать, что Π^d содержит те и только те предложения документа d, в которых встречается, по крайней мере, один из терминов, входящих в T^d, и введем на множестве $T^d \times \Pi^d$ отношение вхождения терминов в предложения, задаваемое *матрицей вхождения* $\Delta^d = \|\delta_{ij}\|$ размерности $m \times n$, где

$$\delta_{ij} = \begin{cases} 1, & \text{если термин } t_i \text{ встречается в предложении } \pi_j, \\ 0, & \text{иначе.} \end{cases} \quad (3)$$

Множество предложений $\Pi_t \subseteq \Pi^d$, в которых встречается термин $t \in T^d$, назовем *носителем* этого термина в документе d:

$$\Pi_t = \{\pi \in \Pi^d : \delta_{\mathrm{ind}(t)\,\mathrm{ind}(\pi)} = 1\}, \quad (4)$$

где $\mathrm{ind}(t)$ – индекс термина t в T^d, $\mathrm{ind}(\pi)$ – индекс предложения π в Π^d, и введем *оператор носителя* терминов как отображение $T^d \to 2^{\Pi^d}$:

$$\forall t \in T^d \quad \mathrm{Supp}\{t\} \equiv \Pi_t . \tag{5}$$

Определение (5) задает оператор носителя терминов только для одноэлементных подмножеств множества T^d. Продолжение оператора носителя терминов на все множество подмножеств T^d может быть задано в виде отображения $2^{T^d} \to 2^{\Pi^d}$:

$$\mathrm{Supp}(T) = \begin{cases} \bigcap\limits_{t \in T} \Pi_t, & \text{если } T \neq \varnothing, \\ \Pi^d, & \text{если } T = \varnothing, \end{cases} \tag{6}$$

где $T \subseteq T^d$.

Множество терминов $T_\pi \subseteq T^d$, которые входят в предложение $\pi \in \Pi^d$, назовем *контентом* этого предложения:

$$T_\pi = \left\{ t \in T^d : \delta_{\mathrm{ind}(t)\,\mathrm{ind}(\pi)} = 1 \right\} \tag{7}$$

и введем *оператор контента* предложений как отображение $\Pi^d \to 2^{T^d}$:

$$\forall \pi \in \Pi^d \quad \mathrm{Cont}\{\pi\} \equiv T_\pi , \tag{8}$$

продолжив его аналогичным образом на все множество подмножеств Π^d в виде отображения $2^{\Pi^d} \to 2^{T^d}$:

$$\mathrm{Cont}(\Pi) = \begin{cases} \bigcap\limits_{\pi \in \Pi} T_\pi, & \text{если } \Pi \neq \varnothing, \\ T^d, & \text{если } \Pi = \varnothing, \end{cases} \tag{9}$$

где $\Pi \subseteq \Pi^d$.

Необходимо отметить, что операторы Supp и Cont не являются взаимообратными. Так, если для некоторого текстового документа, состоящего из двух предложений и содержащего два термина, матрица вхождения имеет вид:

$$\Delta = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \tag{10}$$

то $\mathrm{Cont}\big(\mathrm{Supp}\{t_1\}\big) = \mathrm{Cont}\{\pi_1\} = \{t_1, t_2\} \neq \{t_1\}$. В то же время имеют место следующие обратные включения.

Утверждение 1.

1. $\forall T_1, T_2 \subseteq T^d \quad T_1 \subseteq T_2 \Rightarrow \mathrm{Supp}(T_2) \subseteq \mathrm{Supp}(T_1)$.
2. $\forall \Pi_1, \Pi_2 \subseteq \Pi^d \quad \Pi_1 \subseteq \Pi_2 \Rightarrow \mathrm{Cont}(\Pi_2) \subseteq \mathrm{Cont}(\Pi_1)$.

Доказательство:

1. Если $T_1 \subseteq T_2$, то $T_2 = T_1 \cup T'$, где $T' = T_2 \setminus T_1$. Из определения оператора носителя терминов (6) следует, что $\mathrm{Supp}(T_2) = \mathrm{Supp}(T_1) \cap \mathrm{Supp}(T') \subseteq \mathrm{Supp}(T_1)$. Что и требовалось доказать.

2. Доказывается аналогично. ▥

Замечание. Утверждения, обратные Утверждению 1, в общем случае не верны. Действительно, пусть $T_1 = \{t_1\}$, $T_2 = \{t_1, t_2\}$, тогда из определения оператора носителя терминов (6) и матрицы вхождения (10) имеем $\mathrm{Supp}(T_1) = \mathrm{Supp}(T_2) = \{\pi_1\}$, т.е. имеет место включение $\mathrm{Supp}(T_1) \subseteq \mathrm{Supp}(T_2)$, но при этом $T_1 \subseteq T_2$.

Рассмотрим подмножества терминов документа d, которые устойчиво встречаются в различных его предложениях. Множество пар $C^d = \{\langle T, \Pi \rangle : T \subseteq T^d, \Pi \subseteq \Pi^d\}$, являющихся нетривиальными решениями ($\langle T, \Pi \rangle \neq \langle \varnothing, \varnothing \rangle$) системы уравнений

$$\begin{cases} \Pi = \mathrm{Supp}(T) \\ T = \mathrm{Cont}(\Pi) \end{cases} \qquad (11)$$

назовем *смысловыми контекстами* документа d и обозначим их $[[T, \Pi]]$.

2. Свойства смысловых контекстов

Рассмотрим свойства смысловых контекстов документа d, позволяющие найти общее решение системы уравнений (11).

В дальнейшем для удобства записи обозначим:

$$\begin{aligned} &\mathrm{ContSupp}(T) \equiv \mathrm{Cont}(\mathrm{Supp}(T)), \\ &\mathrm{SuppCont}(\Pi) \equiv \mathrm{Supp}(\mathrm{Cont}(\Pi)). \end{aligned} \qquad (12)$$

Так как операторы Supp и Cont не являются взаимообратными, то, в общем случае, $T \neq \mathrm{ContSupp}(T)$ и $\Pi \neq \mathrm{SuppCont}(\Pi)$, однако имеют место следующие включения.

Утверждение 2.

1. $\forall T \subseteq T^d \quad T \subseteq \mathrm{ContSupp}(T)$.

2. $\forall \Pi \subseteq \Pi^d \quad \Pi \subseteq \mathrm{SuppCont}(\Pi)$.

Доказательство:

1. Покажем, что $\forall T \subseteq T^d$ и $\forall t \in T \quad t \in \mathrm{ContSupp}(T)$.

Возьмем произвольное подмножество терминов $T \subseteq T^d$. Согласно Утверждению 1 для любого термина $t \in T$ имеет место включение $\mathrm{Supp}(T) \subseteq \mathrm{Supp}\{t\}$. Применяя оператор контента к обеим частям включения, получаем, что $\mathrm{ContSupp}\{t\} \subseteq \mathrm{ContSupp}(T)$. Так как произведение $\Delta^d \cdot (\Delta^d)^{\mathrm{T}}$ задает рефлексивное отношение на множестве T^d, то $\{t\} \subseteq \mathrm{ContSupp}\{t\}$, откуда

$\{t\} \subseteq \mathrm{ContSupp}(T)$ и, в силу произвольности выбора $t \in T$, непосредственно следует истинность включения $T \subseteq \mathrm{ContSupp}(T)$.

2. Доказывается аналогично. ˅

Рассмотрим критерии, при которых включения, приведенные в Утверждении 2, выполняются в виде равенств.

Утверждение 3.

1. $T = \mathrm{ContSupp}(T) \quad \Leftrightarrow \quad \exists \Pi \subseteq \Pi^d : T = \mathrm{Cont}(\Pi)$,

2. $\Pi = \mathrm{SuppCont}(\Pi) \quad \Leftrightarrow \quad \exists T \subseteq T^d : \Pi = \mathrm{Supp}(T)$.

Доказательство:

1. Для доказательства прямого утверждения достаточно положить $\Pi = \mathrm{Supp}(T)$.

Докажем обратное утверждение. Из существования $\Pi \subseteq \Pi^d$ такого, что $T = \mathrm{Cont}(\Pi)$ и применения к обеим частям равенства оператора носителя, получаем $\mathrm{Supp}(T) = \mathrm{SuppCont}(\Pi)$. Согласно Утверждению 2 имеем включение $\Pi \subseteq \mathrm{SuppCont}(\Pi)$, откуда следует, что $\Pi \subseteq \mathrm{Supp}(T)$.

Таким образом, носитель множества терминов T можно представить в виде $\mathrm{Supp}(T) = \Pi \cup \Pi'$, где $\Pi' = \mathrm{Supp}(T) \setminus \Pi$. Применяя к обеим частям равенства оператор контента, получаем $\mathrm{ContSupp}(T) = \mathrm{Cont}(\Pi) \cap \mathrm{Cont}(\Pi')$, откуда следует, что $\mathrm{ContSupp}(T) \subseteq \mathrm{Cont}(\Pi) = T$. В то же время, согласно Утверждению 2 имеет место обратное включение $T \subseteq \mathrm{ContSupp}(T)$. Следовательно, $T = \mathrm{ContSupp}(T)$.

2. Доказывается аналогично. ˅

Следствие 1. Из Утверждения 3 непосредственно следует, что $\forall T \subseteq T^d$ и $\forall \Pi \subseteq \Pi^d$ имеют место тождественные равенства:

$$\begin{aligned} &\mathrm{SuppCont}(\mathrm{Supp}(T)) \equiv \mathrm{Supp}(T), \\ &\mathrm{ContSupp}(\mathrm{Cont}(\Pi)) \equiv \mathrm{Cont}(\Pi). \end{aligned} \tag{13}$$

Следствие 2. Из Утверждения 3 следует, что $\forall T \subseteq T^d$ и $\forall \Pi \subseteq \Pi^d$ пары $[[\mathrm{ContSupp}(T), \mathrm{Supp}(T)]]$ и $[[\mathrm{Cont}(\Pi), \mathrm{SuppCont}(\Pi)]]$ являются смысловыми контекстами документа d. Истинность этого следствия непосредственно вытекает из (13):

$$\begin{aligned} &\mathrm{Supp}(\mathrm{ContSupp}(T)) = \mathrm{SuppCont}(\mathrm{Supp}(T)) = \mathrm{Supp}(T), \\ &\mathrm{Cont}(\mathrm{SuppCont}(\Pi)) = \mathrm{ContSupp}(\mathrm{Cont}(\Pi)) = \mathrm{Cont}(\Pi). \end{aligned} \tag{14}$$

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

$$[[T]] \overset{def}{=} [[\mathrm{ContSupp}(T), \mathrm{Supp}(T)]],$$
$$[[\Pi]] \overset{def}{=} [[\mathrm{Cont}(\Pi), \mathrm{SuppCont}(\Pi)]],$$

(15)

где множества терминов T и предложений Π называются *образующими* соответствующих смысловых контекстов. При этом необходимо отметить, что по заданному смысловому контексту нельзя, в общем случае, однозначно определить его образующие.

Рассмотрим свойства контекстов, связанных с пересечением и объединением их образующих.

Утверждение 4. Для $\forall [[T_1, \Pi_1]],\ [[T_2, \Pi_2]] \in C^d$ имеют место равенства:

1. $[[T_1 \cup T_2]] = [[\Pi_1 \cap \Pi_2]]$.

2. $[[T_1 \cap T_2]] = [[\Pi_1 \cup \Pi_2]]$.

Доказательство:

1. Так как $[[T_1, \Pi_1]],\ [[T_2, \Pi_2]] \in C^d$, то $T_1 = \mathrm{Cont}(\Pi_1),\ T_2 = \mathrm{Cont}(\Pi_2)$. Следовательно,

$$\mathrm{ContSupp}(T_1 \cup T_1) = \mathrm{ContSupp}(\mathrm{Cont}(\Pi_1) \cup \mathrm{Cont}(\Pi_2)) = \mathrm{Cont}(\mathrm{SuppCont}(\Pi_1) \cap \mathrm{SuppCont}(\Pi_2)) =$$
$$= \mathrm{Cont}(\Pi_1 \cap \Pi_2).$$

С другой стороны, $\Pi_1 = \mathrm{Supp}(T_1),\ \Pi_2 = \mathrm{Supp}(T_2)$, откуда следует, что $\mathrm{SuppCont}(\Pi_1 \cap \Pi_2) =$

$$= \mathrm{SuppCont}(\mathrm{Supp}(T_1) \cap \mathrm{Supp}(T_2)) = \mathrm{Supp}(\mathrm{ContSupp}(T_1) \cup \mathrm{ContSupp}(T_2)) = \mathrm{Cont}(T_1 \cup T_2).$$

Таким образом, $[[T_1 \cup T_2]] = [[\mathrm{ContSupp}(T_1 \cup T_1), \mathrm{Cont}(T_1 \cup T_2)]] =$

$$= [[\mathrm{Cont}(\Pi_1 \cap \Pi_2), \mathrm{SuppCont}(\Pi_1 \cap \Pi_2)]] = [[\Pi_1 \cap \Pi_2]].$$

2. Доказывается аналогично. ˅

Введем на множестве смысловых контекстов C^d *отношение включения*. Пусть $[[\Pi_1, T_1]], [[\Pi_2, T_2]] \in C^d$, тогда

$$[[\Pi_1, T_1]] \subseteq [[\Pi_2, T_2]] \Leftrightarrow \Pi_1 \subseteq \Pi_2. \qquad (16)$$

Заметим, что отношение включения на множестве смысловых контекстов документа d является рефлексивным, антисимметричным и транзитивным отношением и задает частичный порядок на множестве C^d.

Следствие 3. Из определения (16) с учетом Утверждения 1 получаем: $[[\Pi_1, T_1]] \subseteq [[\Pi_2, T_2]] \Leftrightarrow T_2 \subseteq T_1$.

Объединением смысловых контекстов $[[T_1, \Pi_1]],\ [[T_2, \Pi_2]] \in C^d$ назовем смысловой контекст $[[T_1, \Pi_1]] \vee [[T_2, \Pi_2]] = [[\Pi_1 \cup \Pi_2]]$.

Пересечением смысловых контекстов $[[T_1, \Pi_1]]$, $[[T_2, \Pi_2]] \in C^d$ назовем смысловой контекст $[[T_1, \Pi_1]] \wedge [[T_2, \Pi_2]] = [[\Pi_1 \cap \Pi_2]]$.

Отметим, что операции $\vee$ и $\wedge$ ассоциативны в силу их определения через операции объединения и пересечения множеств и для $\forall C \subseteq C^d$

$$\vee C \stackrel{def}{=} \bigvee_{[[T,\Pi]] \in C} [[T, \Pi]] = [[\bigcup_{[[T,\Pi]] \in C} \Pi]] \text{ и} \qquad (17)$$

$$\wedge C \stackrel{def}{=} \bigwedge_{[[T,\Pi]] \in C} [[T, \Pi]] = [[\bigcap_{[[T,\Pi]] \in C} \Pi]] \ .$$

Обозначим через C_1^d множество базовых смысловых контекстов документа d, состоящее из смысловых контекстов, построенных на одноэлементных подмножествах множества предложений Π^d:

$$C_1^d \stackrel{\Delta}{=} \{ [[\{\pi\}]] : \pi \in \Pi^d \} \cup [[\varnothing]]. \qquad (18)$$

Теорема. Множество смысловых контекстов C^d документа d является замыканием множества базовых смысловых контекстов C_1^d относительно операции $\vee$.

Доказательство.

Пусть $[[T, \Pi]] \in C^d$ – произвольный смысловой контекст. Соответствующее ему подмножество базовых смысловых контекстов $C_1 \subseteq C_1^d$ имеет вид $C_1 = \{ [[\pi]] : \pi \in \Pi \}$. С учетом (18) получаем:

$\vee C_1 = [[\bigcup_{[[T',\Pi']] \in C_1} \Pi]] = [[\Pi]]$. Так как множества T и Π удовлетворяют (11), то $[[\Pi]] = [[T, \Pi]]$.

Таким образом, любой смысловой контекст $[[T, \Pi]] \in C^d$ может быть получен как объединение некоторого подмножества базовых смысловых контекстов. ˅

Замечание. Из доказательства теоремы следует, что построение множества смысловых контекстов может быть осуществлено через множество подмножеств базовых смысловых контекстов $C_1 \subseteq C_1^d$, при чем $|C_1| \le n$. Таким образом, верхняя оценка вычислительной сложности алгоритма построения C^d равна 2^n, что делает неэффективным алгоритм, основанный на переборе всего множества Π^d, для текстовых документов большого объема.

3. Контекстно-ассоциативная модель текста

Предлагаемая модель будет являться развитием ассоциативной модели, предложенной в работе [20].

Пусть $c = [[T, \Pi]] \in C^d$. Назовем первое предложение $\tilde{\pi} \in \Pi$ *порождающим предложением* смыслового контекста c. Оставшуюся часть носителя смыслового контекста $\tilde{\Pi} = \Pi / \{\tilde{\pi}\}$ назовем *областью существования* смыслового контекста c.

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

$$c_\alpha \leftrightarrow c_\beta \quad \Leftrightarrow \quad \tilde{\Pi}_\alpha \cap \tilde{\Pi}_\beta \neq \varnothing. \tag{19}$$

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

$$c_\alpha \leftrightarrow c_{j_1} \leftrightarrow c_{j_2} \leftrightarrow \ \ldots \ \leftrightarrow c_{j_k} \leftrightarrow c_\beta, \tag{20}$$

то *уровнем ассоциативной связи* k между c_α и c_β назовем наименьшую длину такой последовательности. Таким образом, непосредственная ассоциативная связь двух смысловых контекстов – это ассоциативная связь уровня 0.

Вес ассоциативной связи будем рассчитывать через ее уровень k:

$$w(c_\alpha, c_\beta) = 1/2^k. \tag{21}$$

Ассоциативной мощностью уровня l смыслового контекста $c \in C^d$ назовем средневзвешенную сумму весов ассоциативных связей этого смыслового контекста с другими смысловыми контекстами документа d:

$$W^l(c) = \sum_{k=0}^{l} \left(\frac{1}{\left| C_c^k \right|} \sum_{c^* \in C_c^k} w(c, c^*) \right), \tag{22}$$

где C_c^k – множество контекстов, связанных с контекстом c ассоциативной связью уровня k. Уровень l является эвристическим параметром контекстно-ассоциативной модели и выбирается экспериментальным путем.

4. Применение контекстно-ассоциативных моделей в задаче утонения поисковых запросов

Рассмотрим модель ИПС в виде тройки

$$\langle D, Q, f \rangle, \tag{23}$$

где D – конечное множество документов, Q – множество поисковых запросов (в общем случае – бесконечное, заданное языком поисковых запросов L_Q), $f : D \times Q \to [0,1]$ – мера релевантности.

Пусть $q \in Q$ – исходный запрос, заданный пользователем ИПС.

$\tilde{D}_q = F(D, q)$ – отклик ИПС на запрос q – упорядоченная последовательность элементов документов из D, в которой документ с порядковым номером i предшествует элементу с

индексом j, если $f(d_i,q) > f(d_j,q)$, с точностью до перестановок документов с равными оценками релевантности.

Пусть пользователь ИПС, анализируя отклик $\widetilde{D}_q$, сформировал релевантную выборку $\widetilde{D}_q^{rel}$ – конечную подпоследовательность заинтересовавших его документов. В этом случае интегральная оценка качества отклика ИПС имеет вид:

$$\text{quality}(\widetilde{D}_q, \widetilde{D}_q^{\text{rel}}) = \sum_{\forall d \in \widetilde{D}_q^{\text{rel}}} \frac{1}{\text{ind}(d)}. \qquad (24)$$

Задача уточнения запросов состоит в построении нового запроса $q^* \in Q$, такого что $\text{quality}(\widetilde{D}_{q^*}, \widetilde{D}_q^{\text{rel}}) > \text{quality}(\widetilde{D}_q, \widetilde{D}_q^{\text{rel}})$.

(25)

Заметим, что в силу конечности множества документов D для любого $\widetilde{D}_q$ всегда будет существовать оптимальный запрос, улучшить который нельзя. Однако в силу сложности алгоритмов, реализующих различные меры релевантности и бесконечности множества запросов, задача поиска оптимального запроса в общем случае не решена.

Алгоритм уточнения запроса:

1. Для всех документов $d \in \widetilde{D}_q^{\text{rel}}$ строятся их контекстно-ассоциативные модели C^d.

2. Для всех терминов из анализируемых релевантных документов вычисляются весовые коэффициенты:

$$W(t,d) = \frac{1}{\left|C_t^d\right|} \sum_{c \in C_t^d} W^l(c), \qquad (26)$$

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:

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

(27)

4. В уточненный запрос включаются первые *m* терминов, имеющих наибольшие значения весовых коэффициентов. Параметр *m* может устанавливаться пользователем ИПС или выбираться экспериментальным путем.

5. Примеры

Пример 1

Пусть для некоторого текстового документа *d*, состоящего из 6 предложений, включающих 5 терминов, матрица вхождения терминов Δ^d имеет вид:

	π_1	π_2	π_3	π_4	π_5	π_6
t_1	1	1	1	0	1	0
t_2	1	1	0	1	0	0
t_3	0	1	1	0	1	1
t_4	0	0	1	1	1	0
t_5	0	0	0	1	1	1

Примеры носителей и контентов:

Supp $\{t_2, t_3\} = \{\pi_2\}$ и ContSupp $\{t_2, t_3\}$ = Cont $\{\pi_2\} = \{\pi_1, \pi_2, \pi_3\}$

Cont $\{\pi_1, \pi_2, \pi_5\} = \{t_1\}$ и SuppCont $\{\pi_1, \pi_2, \pi_5\}$ = Supp $\{t_1\} = \{\pi_1, \pi_2, \pi_3, \pi_5\}$

Множество базовых смысловых контекстов имеет вид $C_1^d = \{c_i : i = 0, 1, \ldots, 6\}$, где

$c_0 = [[\varnothing]] = [[\varnothing, \{t_1, t_2, \ldots, t_5\}]]$, $c_1 = [[\pi_1]] = [[\{\pi_1, \pi_2\}, \{t_1, t_2\}]]$, $c_2 = [[\pi_2]] = [[\{\pi_2\}, \{t_1, t_2, t_3\}]]$,

$c_3 = [[\pi_3]] = [[\{\pi_3, \pi_5\}, \{t_1, t_3, t_4\}]]$, $c_4 = [[\pi_4]] = [[\{\pi_4\}, \{t_2, t_4, t_5\}]]$, $c_5 = [[\pi_5]] = [[\{\pi_5\}, \{t_1, t_3, t_4, t_5\}]]$,

$c_6 = [[\pi_6]] = [[\{\pi_5, \pi_6\}, \{t_3, t_5\}]]$.

Оставшиеся смысловые контексты порождаются через замыкание множества базовых смысловых контекстов следующим образом:

$c_7 = \vee\{c_3, c_1\} = [[\{\pi_1, \pi_2, \pi_3, \pi_5\}, \{t_1\}]]$, $c_8 = \vee\{c_3, c_2\} = [[\{\pi_2, \pi_3, \pi_5\}, \{t_1, t_3\}]]$,

$c_9 = \vee\{c_4, c_1\} = [[\{\pi_1, \pi_2, \pi_4\}, \{t_2\}]]$, $c_{10} = \vee\{c_4, c_3\} = [[\{\pi_3, \pi_4, \pi_5\}, \{t_4\}]]$,

$c_{11} = \vee\{c_5, c_4\} = [[\{\pi_4, \pi_5\}, \{t_4, t_5\}]]$, $c_{12} = \vee\{c_6, c_1\} = [[\{\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6\}, \varnothing]]$,

$c_{13} = \vee\{c_6, c_2\} = [[\{\pi_2, \pi_3, \pi_5, \pi_6\}, \{t_3\}]]$, $c_{14} = \vee\{c_6, c_4\} = [[\{\pi_4, \pi_5, \pi_6\}, \{t_5\}]]$.

В табл. 1 приведены ассоциативные мощности нулевого уровня смысловых контекстов в порядке их убывания (по столбцам):

Табл. 1

c_{12}	0,929	c_{13}	0,857	c_{10}	0,786	c_3	0,643	c_1	0,429
c_7	0,857	c_{11}	0,786	c_9	0,714	c_5	0,643	c_4	0,357
c_8	0,857	c_{14}	0,786	c_6	0,643	c_2	0,429		

Обобщенные весовые коэффициенты терминов (в порядке убывания):

$t_3 - 0{,}679$, $t_4 - 0{,}643$, $t_5 - 0{,}643$, $t_1 - 0{,}643$, $t_2 - 0{,}482$.

Пример 2

Рассмотрим сегмент теста, состоящий из 5 предложений (табл. 2).

Табл. 2

π_1	*Нормальное приближение для биномиального распределения имеет важное теоретическое и практическое значение в теории вероятностей.*
π_2	*Нормальное приближение сыграло большую роль в развитии теории вероятностей, так как привело к первой предельной теореме.*
π_3	*С современной точки зрения первая предельная теорема является лишь частным случаем*

	центральной предельной теоремы.
π_4	*Нормальное распределение часто называют гауссовским распределением, но оно использовалось в теории вероятностей еще Муавром и Лапласом.*
π_5	*Рассмотрим, как используется нормальное распределение в качестве приближения для биномиального распределения с p=1/2.*

Предложения рассматриваемого сегмента текста включают 36 терминов, матрица вхождения которых показана в виде табл. 3.

Табл. 3

t	Термин	π_1	π_2	π_3	π_4	π_5
1	биномиальный	1				1
2	биномиальное распределение	1				1
3	больший		1			
4	большая роль		1			
5	важный	1				
6	важный теоретический	1				
7	вероятность	1	1		1	
8	значение	1				
9	зрение			1		
10	качество приближения					1
11	нормальный	1	1		1	1
12	нормальное приближение	1	1			
13	нормальное распределение				1	1
14	практический	1				
15	практическое значение	1				
16	предельный		1	1		
17	предельная теорема		1	1		
18	приближение	1	1			1
19	развитие		1			
20	развитие теории		1			
21	распределение	1			1	1
22	роль		1			
23	случай			1		
24	современный			1		
25	современная точка			1		
26	современная точка зрения			1		
27	теорема		1	1		
28	теоретический	1				
29	теория вероятностей	1	1		1	
30	точка			1		
31	точка зрения			1		
32	центральный			1		
33	центральный предельный			1		
34	центральная предельная теорема			1		
35	частный			1		
36	частный случай			1		

Примеры носителей и контентов:

- если T = {"вероятность", "нормальное приближение"}, то Supp(T) = $\{\pi_1, \pi_2\}$ и ContSupp(T) = = {"вероятность", "нормальный", "нормальное приближение", "приближение", "теория вероятностей"}.
- если $\Pi = \{\pi_1, \pi_2, \pi_5\}$, то Cont($\Pi$) = {"нормальный", "приближение"} и SuppCont(Π) = Π.

Множество базовых смысловых контекстов имеет вид $C_1^d = \{c_i : i = 0, 1, \ldots, 5\}$, где

$$c_0 = [[\varnothing]] = [[\varnothing, \{t_1, t_2, \ldots, t_{36}\}]],$$

$$c_1 = [[\pi_1]] = [[\{\pi_1\}, \{t_1, t_2, t_5, \ldots, t_8, t_{11}, t_{12}, t_{14}, t_{15}, t_{18}, t_{21}, t_{28}, t_{29}\}]],$$

$$c_2 = [[\pi_2]] = [[\{\pi_2\}, \{t_3, t_4, t_7, t_{11}, t_{12}, t_{16}, \ldots, t_{20}, t_{22}, t_{27}, t_{29}\}]],$$

$$c_3 = [[\pi_3]] = [[\{\pi_3\}, \{t_9, t_{16}, t_{17}, t_{23}, \ldots, t_{27}, t_{30}, \ldots, , t_{36}\}]],$$

$$c_4 = [[\pi_4]] = [[\{\pi_4\}, \{t_7, t_{11}, t_{13}, t_{21}, t_{29}\}]],$$

$$c_5 = [[\pi_5]] = [[\{\pi_5\}, \{t_1, t_2, t_{10}, t_{11}, t_{13}, t_{18}, t_{21}\}]].$$

Оставшиеся смысловые контексты порождаются через замыкание множества базовых смысловых контекстов следующим образом:

$$c_6 = \vee\{c_2, c_1\} = [[\{\pi_1, \pi_2\}, \{t_7, t_{11}, t_{12}, t_{18}, t_{29}\}]],$$

$$c_7 = \vee\{c_3, c_1\} = [[\{\pi_1, \pi_2, \pi_3, \pi_4, \pi_5\}, \varnothing]],$$

$$c_8 = \vee\{c_3, c_2\} = [[\{\pi_2, \pi_3\}, \{t_{16}, t_{17}, t_{27}\}]],$$

$$c_9 = \vee\{c_4, c_1\} = [[\{\pi_1, \pi_4\}, \{t_7, t_{11}, t_{21}, t_{29}\}]],$$

$$c_{10} = \vee\{c_4, c_2\} = [[\{\pi_1, \pi_2, \pi_4\}, \{t_7, t_{11}, t_{29}\}]],$$

$$c_{11} = \vee\{c_5, c_1\} = [[\{\pi_1, \pi_5\}, \{t_1, t_2, t_{11}, t_{18}, t_{21}\}]],$$

$$c_{12} = \vee\{c_5, c_2\} = [[\{\pi_1, \pi_2, \pi_5\}, \{t_{11}, t_{18}\}]],$$

$$c_{13} = \vee\{c_5, c_4\} = [[\{\pi_4, \pi_5\}, \{t_{11}, t_{13}, t_{21}\}]],$$

$$c_{14} = \vee\{c_9, c_5\} = [[\{\pi_1, \pi_4, \pi_5\}, \{t_{11}, t_{21}\}]],$$

$$c_{15} = \vee\{c_{10}, c_5\} = [[\{\pi_1, \pi_2, \pi_4, \pi_5\}, \{t_{11}\}]].$$

В табл. 4 приведены ассоциативные мощности нулевого уровня смысловых контекстов в порядке их убывания (по столбцам):

Табл. 4

c_7	0,933	c_{10}	0,800	c_{11}	0,667	c_1	0,533	c_4	0,400
c_{15}	0,867	c_{14}	0,733	c_9	0,667	c_8	0,467	c_5	0,400
c_{12}	0,800	c_6	0,667	c_{13}	0,600	c_2	0,400	c_3	0,133

Обобщенные весовые коэффициенты терминов (в порядке убывания) в случае применения контекстно-ассоциативной модели 1-го уровня приведены в табл. 5,а.

Табл. 5,а

$W(t)$	t
0.398	**нормальный**
0.366	**вероятность**
0.366	**приближение**
0.366	теория вероятностей
0.366	**распределение**
0.281	**нормальное приближение**
0.281	**биномиальное распределение**
0.281	**биномиальный**
0.281	**нормальное распределение**
0.203	центральная предельная теорема
0.203	теорема
0.203	предельный
0.203	практический
0.203	значение
0.188	практическое значение
0.188	центральный
0.188	теоретический

Табл. 5,б

$W(t)$	t
0.692	предельная теорема
0.692	теорема
0.692	предельный
0.590	**нормальный**
0.590	**приближение**

		0.590 **вероятность**

0.590	теория вероятностей
0.590	**нормальное приближение**

0.385	**биномиальное распределение**
0.385	**биномиальный**
0.385	точка
0.385	**нормальное распределение**
0.385	современная точка зрения

0.234	**распределение**
0.234	центральная предельная теорема
0.234	частный случай
0.234	центральный

Для сравнения в табл. 5,б представлены весовые коэффициенты терминов, полученные с использованием ассоциативной модели, предложенной в работе [20] (ключевые термины выделены жирным шрифтом). Видно, что в случае применения контекстно-ассоциативной модели ключевые термины располагаются ближе к началу списка. Вследствие этого, в случае применения контекстно-ассоциативной модели, нет необходимости вводить какие-либо специальные признаки, выделяющие ключевые термины среди всех терминов анализируемого текста.

Пример 3

Работу алгоритма уточнения запросов можно проиллюстрировать примером поиска текстов, посвященных нормальному приближению биномиального распределения.

1. Информационная потребность: найти информацию по нормальному приближению биномиального распределения. Пусть известно, что в тестовой коллекции документов информационной потребности соответствуют 11 релевантных документов.

2. Исходный запрос состоит из 1 термина: q = "приближение биномиального распределения". Отклик ИПС на исходный запрос имеет вид (релевантные документы отмечены серым фоном):

$\widetilde{D}_q =$

d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	…

Пользователь ИПС, анализируя отклик, сформировал релевантную выборку $\widetilde{D}_q^{\text{rel}} = \{d_2, d_5\}$, при этом качество поиска по исходному запросу $\text{quality}(\widetilde{D}_q, \widetilde{D}_q^{\text{rel}}) = 1/2 + 1/5 = 0{,}7$, точность на уровне первых 10 документов $\text{prec}_{10}(q) = 4/10 = 0{,}4$, полнота поиска $\text{recall}_{10}(q) = 4/11 \approx 0{,}367$.

3. Уточненный запрос включает 2 термина: q^* = "нормальное приближение" & "биномиальное распределение". Отклик ИПС на уточненный запрос:

$\widetilde{D}_{q^*} =$

d_5	d_7	d_2	d_{21}	d_9	d_{15}	d_{17}	d_8	d_4	d_{11}	…

Качество поиска по уточненному запросу увеличилось: $\text{quality}(\widetilde{D}_{q^*}, \widetilde{D}_q^{\text{rel}}) = 1 + 1/3 \approx 1{,}333$, при этом точность и полнота поиска не изменились в силу того, что не изменилось число релевантных документов среди первых 10 в отклике ИПС.

Заключение

В заключение хотелось бы отметить, что практические эксперименты, основанные на случайных выборках поисковых запросов, показали применимость предлагаемого метода уточнения поисковых запросов при поиске в специально организованных тестовых коллекциях документов из областей знаний с устоявшейся терминологией (технической документации,

коллекциям учебных пособий и справочным материалам), а также в коллекциях текстов небольшого объема (архивам новостей).

В дальнейшем, для уточнения эвристических параметров алгоритмов и исследования зависимости работы метода от наборов данных, планируется проведение экспериментов с реальными коллекциями документов, в частности, с коллекцией новостей сайта Lenta.Ru, коллекцией технической документации сайта CITForum.Ru и электронной библиотекой кафедры математической кибернетики МАИ.

Список литературы

1. Ермаков А.Е. Полнотекстовый поиск: проблемы и их решение.// Мир ПК. №5, 2000. – http://www.osp.ru/pcworld/2001/05/064.htm (15.05.2001)
2. Furnas G.W., Landauer T.K., Gomez L.M., Dumais S.T. The vocabulary problem in human-system communication. Communications of the ACM, 30(11): 964-971, November 1987.
3. Ашманов И.С. Национальные особенности поисковых систем.// Компьютер в школе. №1, 2000. – http://www.osp.ru/school/2000/01/012.htm (19.01.2000)
4. Sparck-Jones K., Jackson D.M. The use of automatically-obtained keyword classifications for information retrieval. Information Processing and Management, 5:175-201, 1970.
5. Crouch C.J., Yang B. Experiments in automatic statistical thesaurus construction, In Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, 1992, pp. 77-88.
6. Добрынин В.Ю., Некрестьянов И.С. Расширение запросов с помощью вероятностного латентного семантического индексирования. Труды 3-й Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", Петрозаводск, Россия, сентябрь 2001. – с. 151-155.
7. Jing Y., Croft W.B. An association thesaurus for information retrieval. In Proceedings of RIAO-94, 1994, pp. 146-160.
8. Qiu Y., Frei H.P. Concept based query expansion. In Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, 1993, pp. 160-169.
9. Attar R., Fraenkel T.S. Local feedback in full-text retrieval systems. Journal of the Association for Computing Machinery, 24(3), July 1977, pp. 397-417.
10. Croft W.B., Xu J. Query expansion using local and global document analysis. In Proc. of the SIGIR'96, 1996, pp. 4-11.
11. Xu J., Croft W.B. Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems, 18(1), 2000, pp. 79-112.

12. Croft W.B., Harper D.J. Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35:285-295, 1979.

13. Rocchio J.J. Relevance feedback in information retrieval. In G. Salton (Ed.), The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice Hall, 1971, pp. 313-323.

14. Salton G., Buckley C. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4), 1990, pp. 182-188.

15. Allan J. Relevance Feedback With Too Much Data. Research and Development in Information Retrieval, 1995, pp. 337-343.

16. Grootjen F.A., Th.P. van der Weide. Conceptual Query Expansion. Technical Report NIII-R0406, Nijmegen Institute for Information and Computing Sciences, University of Nijmegen, Nijmegen, The Netherlands, EU, 2004.

17. Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval. ACM Press, 1999.

18. Cui H., Wen J.-R., Nie J.-Y., Ma W.-Y. Probabilistic query expansion using query logs. In Proceedings of the eleventh international conference on World Wide Web (2002), ACM Press, pp. 325-332.

19. Silverstein C., Henzinger M., Marais H., Moricz M. Analysis of a very large AltaVista query log. Technical Report 1998-014, COMPAQ System Research Center, October 1998.

20. Чанышев О.Г. Ассоциативная модель естественного текста.// Вестник ОмГУ. – 1997, №4. – с. 17-20.

Беляев Дмитрий Владимирович, аспирант кафедры математической кибернетики Московского авиационного института (государственного технического университета);
e-mail: belyaev@oviont.ru, dvbelyaev@rambler.ru; контактный телефон: (095) 211-3324, 158-4811.