Л.А. Панкова, В.А. Пронина

Семантический текстовый поиск, основанный на теории нечетких множеств

Аннотация. Предлагаются модели текстового поиска на основе теории нечетких множеств, которые используют принцип обобщения для перехода от отношения между понятиями к отношениям между документом и запросом, а именно, от связанности понятий к релевантности документа запросу. Дан пример, показывающий результаты моделирования.

Ключевые слова: текстовый поиск, связанность, релевантность, теория нечетких множеств, нечеткое отношение, принцип обобщения.

Введение

В данной работе рассматривается семантический поиск по запросу в коллекции научных документов на основании содержимого этих документов. Модель текстового поиска включает модель поискового запроса, модель документа и модель релевантности (соответствия) документа запросу. В работе исследованы модели запроса и документа как наборы понятий (терминов) онтологии предметной области коллекции с коэффициентами (весами) от 0 до 1, отражающими важность понятий для описания содержания. В запросе назначенные пользователем веса определяют его информационную потребность. В документах в автоматическом процессе концептуального индексирования [1] распознаются термины понятий и связи между ними, а также определяются веса понятий. Существуют различные методы вычисления весов понятий: с использованием частоты встречаемости, мест встречаемости и др. Релевантность (семантическое соответствие) документа запросу в рассматриваемых моделях определяется с использованием отношения семантической связанности понятий. Семантическая связанность терминов может вычисляться по онтологии предметной области данной коллекции или с использованием статистических методов, а может задаваться экспертом.

Текстовый поиск имеет дело с нечеткой априорной информацией, что не принимается в расчет в большинстве существующих четких моделей [2]. Теория нечетких множеств дает средства обращения с нечеткой информацией. В существующих работах по информационному поиску теория нечетких множеств применяется в основном для представления онтологии и реализации более гибких способов формулирования запросов. На практике применяемые модели информационного поиска по-прежнему основаны на других подходах. Отчасти это можно объяснить тем, что необходимы экспериментальные доказательства, чтобы продемонстрировать, действительно ли нечеткие модели в состоянии превзойти современные подходы.

В данной работе понятия текстового поиска интерпретируются в терминах теории нечетких множеств и рассматриваются модели текстового поиска, использующие теорию нечетких множеств [3]. Предлагаемые модели используют принцип обобщения – универсальный принцип теории нечетких множеств [4] – для перехода от отношений между понятиями к отношениям между документами и запросами, от связанности понятий к релевантности документов и запросов.

В первом разделе даются основные понятия текстового поиска в терминах теории нечетких

множеств, во втором - предложены модели текстового поиска, основанные на теории нечетких множеств. В третьем разделе дан модельный пример ранжирования коллекции документов по релевантности запросу, вычисленной предложенными методами.

1. Понятия текстового поиска в терминах теории нечетких множеств

Интерпретируем понятия текстового поиска в терминах теории нечетких множеств. Пусть D – конечное множество документов коллекции, C – конечное множество понятий предметной области коллекции, Q – конечное множество запросов.

1. Множество концептуальных индексов документов можно представить как нечеткое бинарное индексирующее отношение I:

$$I = \{\mu_I(d,c)/(d,c) \mid d \in D; c \in C\},$$

где μ_I: $D \times C \rightarrow [0, 1]$ – функция принадлежности, обозначающая для каждой пары (d, c) степень принадлежности понятия c документу d (вес понятия в концептуальном индексе).

Индексирующее отношение I индуцирует множества I_d (концептуальные индексы) как нечеткие множества на множестве понятий:

$$I_d = \{\mu_{I_d}(c)/c \mid c \in C, \mu_{I_d}(c) = \mu_I(d,c)\},$$

где $\mu_{I_d}(c)$ – вес понятия в концептуальном индексе документа.

2. Множество концептуальных индексов запросов можно представить как нечеткое бинарное индексирующее отношение:

$$U = \{\mu_U(q,c)/(q,c) \mid q \in Q; c \in C\},$$

где $\mu_U(q,c)$ – функция принадлежности, обозначающая для каждой пары (q, c) степень информационной потребности понятия c в запросе q (вес понятия в концептуальном индексе запроса).

Запрос q представляется как нечеткое множество понятий:

$$I_q = \{\mu_{I_q}(c)/c \mid c \in C, \mu_{I_q}(c) = \mu_U(q,c)\}.$$

3. Отношение семантической связанности понятий S можно представить как нечеткое рефлексивное отношение на $C \times C$ с функцией принадлежности $\mu_S(c_i, c_j) = s(c_i, c_j)$, где $s(c_i, c_j) \in [0,1]$ – семантическая связанность понятий c_i и c_j:

$$S = \{\mu_S(c_i,c_j)/(c_i,c_j) \mid c_i, c_j \in C\}.$$

2. Нечеткие модели текстового поиска, основанные на принципе обобщения

Предлагаемые модели используют принцип обобщения и интуитивно просты. Принцип обобщения – это универсальный принцип теории нечетких множеств. В предлагаемых моделях принцип обобщения используется для перехода от отношения на понятиях к отношению на документах и запросах, а именно, от связанности понятий к релевантности документов и запросов.

На первом этапе отношение связанности на понятиях обобщается, чтобы получить нечеткое отношение связанности S' нечетких запросов с одним понятием:

$$\mu_{S'}(I_q, c_j) = \max\{\min_{c_i \in C}\{\mu_{I_q}(c_i), \mu_S(c_i, c_j)\}\}.$$

Затем принцип обобщения используется еще раз. При этом нечеткое отношение связанности S' нечетких запросов с понятием обобщается, чтобы получить обобщенное нечеткое отношение связанности S'' на $\{I_q\} \times \{I_d\}$ с функцией принадлежности $\mu_{S''}(I_q, I_d)$, которую будем называть ***обобщенной связанностью запроса и документа***:

$$\mu_{S''}(I_q, I_d) = \max\{\min_{c_j \in C}\{\mu_{I_d}(c_j), \mu_{S'}(I_q, c_j)\}\} =$$

$$= \max\{\min_{c_j \in C}\{\mu_{I_d}(c_j), \max \min_{c_i \in C}\{\mu_{I_q}(c_i), \mu_S(c_i, c_j)\}\}\}.$$

Эта формула преобразуется к виду:

$$\mu_{S''}(I_q, I_d) = \max\{\min_{c_i, c_j \in C}\{\mu_{I_q}(c_j), \mu_{I_d}(c_i), \mu_S(c_i, c_j)\}\}.$$

Таким образом, обобщенное нечеткое отношение связанности запросов и документов S'' имеет вид:

$$S'' = \{\mu_{S''}(I_q, I_d)/(I_q, I_d) \mid I_q \in \{I_q\}, I_d \in \{I_d\}\}.$$

Рис. 1 – Рис. 3 иллюстрируют применение принципа обобщения.

Если I_q и I_d – одноэлементные нечеткие множества, то S'' интерпретируется как обобщенное нечеткое отношение связанности двух

нечетких понятий с функцией принадлежности (максимум исчезает, так как рассматривается одна пара понятий), которое будем называть ***обобщенной связанностью двух понятий***:

$$\mu_{S''}(\tilde{c}_i,\tilde{c}_j)=\min\{\mu_{I_q}(c_i),\mu_{I_d}(c_j),\mu_S(c_i,c_j)\}.$$

Заметим, что обобщенную связанность I_q и I_d можно представить как максимальное значение обобщенных связанностей всех пар нечетких понятий запроса и документа:

$$\mu_{S''}(I_q,I_d)=\max_{\tilde{c}_i\subset I_q,\tilde{c}_j\subset I_d}\{\mu_{S''}(\tilde{c}_i,\tilde{c}_j)\}.$$

Обобщенную связанность запроса и документа можно использовать для определения релевантности – семантического соответствия – документа запросу. Обобщенная связанность характеризует максимальную связанность пар понятий запроса и документа. Однако максимальная связанность может достигаться на небольшом числе пар понятий, а остальные пары понятий могут иметь незначительную связанность. Такой документ может быть предпочтен документу с несколько меньшей обобщенной связанностью, но достигаемой на большем числе пар. Введем параметр ширины связанности запроса и документа, использование которого позволит избежать этого эффекта.

Ширина связанности запроса и документа N (q,d) определяет число пар понятий запроса и документа, для которых обобщенная связанность принадлежит заданному полуинтервалу:

$$N(q,d)=|\{(c_i,c_j)\,|\,\mu_{S''}(\tilde{c}_i,\tilde{c}_j)\in(\delta_1,\delta_2]\}|,$$

$$\tilde{c}_i\subset I_q,\tilde{c}_j\subset I_d,\ \ \delta_1<\delta_2,\ \ \delta_1,\delta_2\in(0,\mu_{S''}(I_q,I_d)].$$

Введенные понятия используются для определения релевантности (семантического соответствия) документа запросу и ранжирования документов. Предлагается два способа ранжирования документов.

1. Релевантность документа запросу $R(q,d)$ вычисляется как обобщенная связанность запроса и документа $\mu_{S''}(I_q,I_d)$. Определяется α-срез D_α – множество документов D_α, релевантных запросу с релевантностью большей α. Документы из D_α ранжируются по значению релевантности:

$$R(q,d)=\mu_{S''}(I_q,I_d)=$$

$$=\max_{\tilde{c}_i\subset I_q,\tilde{c}_j\subset I_d}\{\mu_{S''}(\tilde{c}_i,\tilde{c}_j)\},\ d\in D_\alpha.$$

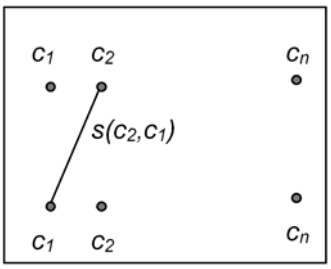

Рис. 1. Семантическая связанность понятий

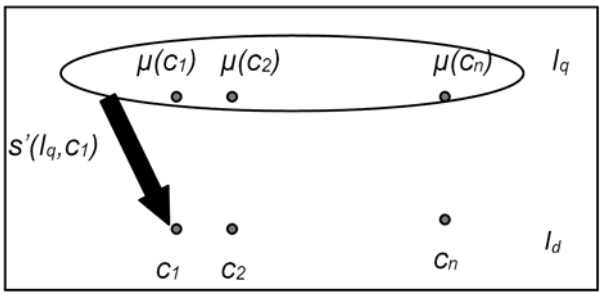

Рис. 2. Семантическая связанность запроса и понятия

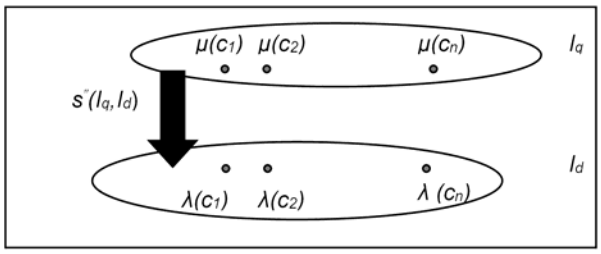

Рис. 3. Семантическая связанность запроса и документа

2. Для ранжирования документов используются обобщенная связанность $\mu_{S''}(I_q,I_d)$ и ширина связанности $N(q,d)$ по следующему алгоритму:

- для каждого документа вычисляется обобщенная связанность запроса и данного документа $\mu_{S''}(I_q,I_d)$;
- определяется α-срез $D_{\alpha;}$;
- на полуинтервале $(\alpha,1)$ вводится лингвистическая переменная «релевантность». Множество документов D_α разбивается на классы эквивалентности D_α^i в соответствии со значениями «релевантности» (например, «сильная», «умеренная», «слабая»):

$$D_\alpha^i=\{d\,|\,\delta_{i1}<\mu_{S''}(I_q,I_d)\le\delta_{i2},i=1,...,k\}$$

где k – число значений лингвистической переменной «релевантность».

Классы эквивалентности ранжируются по значению лингвистической переменной «релевантность»:

- для каждого документа $d \in D_\alpha^i$ вычисляется $N^i(q,d)$:

$$N^i(q,d) = \\ = |\{(c_i, c_j) \mid \mu_{S''}(\widetilde{c}_i, \widetilde{c}_j) \in (\delta_{i1}, \delta_{i2}], \widetilde{c}_i \subset I_q, \widetilde{c}_j \subset I_d\}|.$$

- внутри каждого класса эквивалентности документы ранжируются по $N^i(q,d)$.

3. Пример

Пусть множество C состоит из следующих понятий:

c_1 = *Fuzzy logic;*
c_2 = *Fuzzy relation equations;*
c_3 = *Fuzzy modus ponens;*
c_4 = *Approximate reasoning;*
c_5 = *Max-min composition;*
c_6 = *Fuzzy implication.*

Отношение семантической связанности понятий S (необходимый для вычислений фрагмент) задано матрицей:

$$S = \begin{bmatrix} & c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \\ c_1 & 1 & 0{,}2 & 1 & 1 & 0{,}5 & 1 \\ c_2 & 0{,}2 & 1 & 0{,}1 & 0{,}7 & 0{,}9 & 0 \\ c_3 & 1 & 0{,}4 & 1 & 0{,}9 & 0{,}3 & 1 \end{bmatrix}.$$

Запрос I_q представлен вектором:

$$I_q = \begin{bmatrix} c_1 & c_2 & c_3 \\ 1 & 0{,}4 & 0{,}1 \end{bmatrix}.$$

Дано множество документов $D = \{d_1, \ldots, d_{10}\}$. Индексирующее отношение задано матрицей I (i-й столбец – это концептуальный индекс i-го документа):

$$I = \begin{bmatrix} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10} \\ c_1 & 0{,}2 & 0 & 0{,}1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ c_2 & 1 & 0 & 0 & 0{,}3 & 0 & 0{,}4 & 0 & 0 & 1 & 0 \\ c_3 & 0 & 0 & 0{,}8 & 0 & 0{,}4 & 0 & 1 & 0 & 0 & 0 \\ c_4 & 0 & 0{,}1 & 0 & 0 & 0 & 0 & 0 & 0{,}9 & 0{,}7 & 0{,}5 \\ c_5 & 1 & 0 & 0{,}5 & 0 & 0 & 0{,}6 & 0 & 0 & 0 & 0 \\ c_6 & 0 & 0{,}1 & 0 & 0 & 0{,}2 & 0 & 1 & 0 & 0 & 0{,}5 \end{bmatrix}.$$

3.1. Ранжирование по обобщенной связанности запроса и документа

Релевантности документов запросу вычисляются как обобщенные связанности запроса и документов:

$$\overline{R}(q,d) = \begin{bmatrix} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10} \\ 0{,}5 & 1 & 1 & 0{,}3 & 0{,}4 & 0{,}5 & 1 & 0{,}9 & 0{,}7 & 0{,}5 \end{bmatrix}.$$

Определяется α-срез при $\alpha = 0,5$:

$$\overline{R}_{0,5}(q,d) = \begin{bmatrix} d_2 & d_3 & d_7 & d_8 & d_9 \\ 1 & 1 & 1 & 0{,}9 & 0{,}7 \end{bmatrix}.$$

Документы упорядочиваются по значению $R_{0,5}(q, d)$:

$R_{0,5}(q, d) = 1$: $\{d_2, d_3, d_7\}$,
$R_{0,5}(q, d) = 0{,}9$: $\{d_8\}$,
$R_{0,5}(q, d) = 0{,}7$: $\{d_9\}$.

3.2. Ранжирование по обобщенной связанности и по ширине связанности

Вычисляется α-срез при $\alpha = 0{,}5$:

$$\overline{R}_{0,5}(q,d) = \begin{bmatrix} d_2 & d_3 & d_7 & d_8 & d_9 \\ 1 & 1 & 1 & 0{,}9 & 0{,}7 \end{bmatrix}.$$

Полуинтервал (0,5;1) делится на два полуинтервала (0,5;0,9) и (9, 1), соответствующих значениям лингвистической переменной: «умеренная» и «сильная» релевантность. Множество документов $D_{0.5}$ разбивается на два класса эквивалентности: $\{d_8, d_9\}$ и $\{d_2, d_3, d_7\}$.

Внутри каждого класса документы ранжируются по $N(q,d)$. Итоговая ранжировка имеет вид:

«сильная релевантность», $N(q,d)=3$: $\{d_7\}$,
«сильная релевантность», $N(q,d)=2$: $\{d_2\}$,
«сильная релевантность», $N(q,d)=1$: $\{d_3\}$,
«умеренная релевантность», $N(q,d)=1$: $\{d_8, d_9\}$.

Заметим, что оба способа ранжирования заданной коллекции документов дают близкие результаты и соответствуют неформальному анализу исходных данных: запроса, концептуальных индексов документов и отношения семантической близости понятий.

Заключение

Понятия текстового поиска интерпретируются в терминах теории нечетких множеств.

Предлагаются модели текстового поиска в рамках теории нечетких множеств. Проверена адекватность предложенных методов текстового поиска на модельном примере. Для сравнения моделей текстового поиска, основанных на теории нечетких множеств, с их четкими аналогами следует провести экспериментальную проверку на реальных коллекциях научно-технических текстов.

Литература

1. Соловьев В.Д., Добров Б.В., Иванов В.В., Лукашевич Н.В. Онтологии и тезаурусы: Учебное пособие. Казань, Москва: Казанский государственный университет, МГУ им. М.В. Ломоносова, 2006.
2. Панкова Л.А., Пронина В.А., Крюков К.В. Онтологические модели поиска экспертов в системах управления знаниями научных организаций // Проблемы управления. – 2011. – № 6. – С. 52–60.
3. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. СПб.: Питер. 2000.
4. Орловский С.А. Проблемы принятия решений при нечеткой исходной информации. М: Наука. 1981.

Панкова Людмила Александровна. Старший научный сотрудник ИПУ им. В.А. Трапезникова РАН. Окончила МГУ им. М.В. Ломоносова в 1969 году. Кандидат технических наук. Автор более 80 печатных работ. Область научных интересов: искусственный интеллект, инженерия знаний, интеллектуальные технологии, мягкие вычисления.
E-mail: pankova@ipu.ru

Пронина Валерия Александровна. Старший научный сотрудник ИПУ им. В.А. Трапезникова РАН. Окончила МЭИ (Технический университет) в 1965 году, МГУ им. В.М. Ломоносова в 1971 году. Кандидат технических наук. Автор более 70 печатных работ. Область научных интересов: искусственный интеллект, инженерия знаний, интеллектуальные технологии, мягкие вычисления. E-mail: pron@ipu.ru