УДК 001(06)

МОДЕЛИ ТЕКСТОВОГО ПОИСКА НА ОСНОВЕ ТЕОРИИ НЕЧЕТКИХ МНОЖЕСТВ

Л.А. Панкова, В.А. Пронина

Понятия текстового поиска интерпретированы в терминах теории нечетких множеств. Предложены модели текстового поиска на основе теории нечетких множеств. Показано, что три модели (в том числе две предложенные) дают одну формулу вычисления релевантности документа запросу.

Ключевые слова: текстовый поиск, семантическая связанность, нечеткое множество, нечеткое отношение релевантности, принцип обобщения.

ВВЕДЕНИЕ

Работа посвящена семантическому поиску текстовых документов в коллекции научных документов по их содержимому. Модель текстового поиска на основе онтологии включает в себя модель поискового запроса, модель документа и модель релевантности (соответствия) документа запросу. Онтология предметной области представляет собой формализованное описание терминологии предметной области, отражающее синонимию и семантическую связанность понятий.

В работе рассматриваются модели запроса и документа как наборы понятий (терминов[1]) онтологии предметной области коллекции (словаря терминов) 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 формально определяется с использованием отношения семантической связанности (relatedness) понятий. Семантическая связанность понятий может вычисляться формальным образом по онтологии предметной области данной коллекции или с помощью статистических методов, а может задаваться экспертом. В последнем случае оценка семантической связанности — это оценка возможности с точки зрения эксперта (например, по лингвистической шкале) того, что если в тексте содержится понятие c_i, то в нем будет содержаться и понятие c_j.

Отметим, что текстовый поиск имеет дело с нечеткой априорной информацией, что не принимается в расчет в большинстве существующих четких (crisp) моделей (см., например, обзор [2]). Теория нечетких множеств дает средства обращения с нечеткостями. В данной работе рассматриваются модели текстового поиска, основанные на теории нечетких множеств.

1. ПОНЯТИЯ ТЕКСТОВОГО ПОИСКА В ТЕРМИНАХ ТЕОРИИ НЕЧЕТКИХ МНОЖЕСТВ

Интерпретируем понятия текстового поиска в терминах теории нечетких множеств [3, 4].

Пусть D — конечное множество документов коллекции, C — конечное множество понятий предметной области коллекции, Q — конечное множество запросов.

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

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

где $\mu_I : D \times C \to [0, 1]$ — 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 (концептуаль-

[1] В коллекции научных документов понятия и термины чаще всего не различаются.

ные индексы) как нечеткие множества на множестве понятий:

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

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

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

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

где μ_U: $Q \times C \to [0, 1]$ — 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:

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

1.3. 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:

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

1.4. С целью повышения эффективности текстового поиска вводится понятие расширенного запроса: исходный запрос, дополненный семантически связанными понятиями.

1.4.1. Расширенный запрос представляется нечетким множеством I_q^*, включающим в себя исходный запрос I_q, дополненный семантически связанными понятиями со значениями связанности больше заданного порога, $I_q \subset I_q^*$.

1.4.2. Расширенный запрос представляется как образ нечеткого множества I_q при нечетком отображении (отношении) S — нечеткое множество I_q^*:

$$I_q^* = \sum_i \mu_{I_q^*}(c_i)/c_i.$$

По принципу обобщения[2] функция принадлежности множества I_q^* имеет вид:

$$\mu_{I_q^*}(c_j) = \max\left\{\min_{c_i \in C}\{\mu_{I_q}(c_j), \mu_S(c_i, c_j)\}\right\}.$$

1.4.3. Расширенный запрос представляется как результат композиции двух нечетких отношений I_q и S, если рассматривать I_q как нечеткое унарное отношение на множестве C:

$$I_q^* = I_q \circ S,$$

$$I_q^* = \sum_i \mu_{I_q^*}(c_j)/c_j.$$

Композиция нечетких отношений определяется разными способами. Максиминная композиция нечетких отношений I_q и S на C определяется функцией принадлежности вида

$$\mu_{I_q^*}(c_j) = \max\left\{\min_{c_i \in C}\{\mu_{I_q}(c_i), \mu_S(c_i, c_j)\}\right\},$$

минимаксная композиция — функцией принадлежности вида

$$\mu_{I_q^*}(c_j) = \min\left\{\max_{c_i \in C}\{\mu_{I_q}(c_i), \mu_S(c_i, c_j)\}\right\},$$

максимультипликативная композиция — функцией принадлежности вида

$$\mu_{I_q^*}(c_j) = \max_{c_i \in C}\{\mu_{I_q}(c_i)\mu_S(c_i, c_j)\}.$$

1.5. Релевантность документа запросу определяется различными способами.

1.5.1. Релевантность документов запросу определяется как близость между двумя нечеткими множествами. Близость между двумя нечеткими множествами вычисляется различными способами [5], например как обобщение мер близости между двумя четкими множествами: Хемминга, Эвклида, Дайса и др.

1.5.2. Релевантность документов запросу определяется как образ нечеткого множества I_q на четкое множество D при нечетком индексирующем отображении (отношении) I:

$$R_q = \sum_k \mu_{R_q}(d_k)/d_k.$$

По принципу обобщения функция принадлежности множества R_q имеет вид:

$$\mu_{R_q}(d_k) = \max\left\{\min_{c_i \in C}\{\mu_{I_q}(c_i), \mu_I(d_k, c_i)\}\right\}.$$

Функция $\mu_{R_q}(d_k)$ определяет релевантность документа d_k запросу q.

1.5.3. Релевантность документов запросу определяется как результат композиции двух нечетких отношений I_q и I, если рассматривать I_q как нечеткое унарное отношение на множестве C:

$$R_q = I_q \circ I,$$

$$R_q = \sum_k \mu_{R_q}(d_k)/d_k,$$

[2] Принцип обобщения (generalization principle) — это способ расширения области определения (области значений) отображений (отношений) на класс нечетких множеств.

где $\mu_{R_q}(d_k)$ определяет релевантность документа d_k запросу q, вычисляется в зависимости от выбранной композиции.

2. БЛИЗКИЕ РАБОТЫ

В работе [6] запрос и документ представлены нечеткими множествами I_q и I_d соответственно. Запрос не расширяется и связанность понятий не учитывается. Релевантность документа запросу определяется как близость между нечеткими множествами I_q и I_d по нечеткой мере:

$$R(d, q) = \sum_{c \in C} \min(\mu_{I_d}(c), \mu_{I_q}(c)) / \sum_{c \in C} \mu_{I_q}(c).$$

В работе [7] предлагается учитывать отношения связанности понятий, заданные нечетким отношением семантической связанности S, и расширить запрос, применяя композицию отношений I_q и S (см. пп. 1.4.2):

$$I_q^* = I_q \circ S,$$

$$I_q^* = \sum_i \mu_{I_q^*}(c_i)/c_i.$$

Далее релевантность документа запросу определяется как результат композиции двух нечетких отношений I_q^* и I, при этом I_q^* рассматривается как нечеткое унарное отношение на множестве C (см. пп. 1.5.3):

$$R_q = I_q^* \circ I,$$

$$R_q = \sum_k \mu_{R_q}(d_k)/d_k.$$

Для фильтрации документов со значениями оценок близости, больших $\alpha \in [0, 1]$, используется α-срез нечеткого множества R_q:

$$R_{q(\alpha)} = \{\mu_{R_q}(d)/d | d \in D;\ \mu_{R_q}(d) > \alpha\}.$$

3. МОДЕЛИ ТЕКСТОВОГО ПОИСКА, ОСНОВАННЫЕ НА ТЕОРИИ НЕЧЕТКИХ МНОЖЕСТВ

3.1. Модель текстового поиска, основанная на обобщении нечеткого отношения

В предлагаемой модели используется принцип обобщения — универсальный принцип теории нечетких множеств теории — для перехода от отношений между понятиями к отношениям между документами и запросами, от связанности понятий к релевантности документов запросам.

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

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

Затем принцип обобщения используется еще раз. При этом нечеткое отношение связанности S' нечетких запросов с одним понятием обобщается, чтобы получить нечеткое отношение релевантности нечетким запросам нечетких документов R:

$$\mu_R(I_q, I_d) = \max\left\{\min_{c_j \in C}\{\mu_{I_q}(c_j), \mu_S(I_q, c_j)\}\right\} =$$

$$= \max\left\{\min_{c_j \in C}\left\{\mu_{I_d}(c_j), \max\min_{c_i \in C}\{\mu_{I_q}(c_i), \mu_S(c_i, c_j)\}\right\}\right\}. \quad (1)$$

Можно показать, что формула (1) преобразуется к виду

$$\mu_R(I_q, I_d) =$$

$$= \max\left\{\min_{c_i, c_j \in C}\{\mu_{I_q}(c_j), \mu_{I_d}(c_i), \mu_S(c_i, c_j)\}\right\}. \quad (2)$$

Таким образом, релевантность документа I_d запросу I_q вычисляется по формуле (2).

3.2. Модель текстового поиска с расширением запроса, основанная на максиминной близости между нечеткими множествами

В предлагаемой модели используется расширение запроса максиминной композицией нечетких отношений I_q и S (см. пп. 1.4.3):

$$\mu_{I_q^*}(c_j) = \max\left\{\min_{c_i \in C}\{\mu_{I_q}(c_i), \mu_S(c_i, c_j)\}\right\}.$$

Функция $\mu_{I_q^*}(c_j)$ определяет степень принадлежности понятия c_j расширенному запросу I_q^*.

Релевантность документа запросу определяется как близость между двумя нечеткими множествами: расширенным запросом и документом.

Определим близость между нечеткими множествами A и B на множестве X следующим образом:

$$S(A, B) = \max\left\{\min_{x \in X}\{\mu_A(x), \mu_B(x)\}\right\}.$$

Тогда релевантность документа запросу с использованием введенной близости будет вычисляться по формуле:

$$R(I_q, I_d) = S(I_q, I_d) = \max\min_{c_i \in C}\left\{\mu_{I_q^*}(c_i), \mu_{I_q}(c_i)\right\} =$$

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

что эквивалентно формуле (2).

3.3. Анализ моделей

Модель релевантности из работы [7] можно представить как последовательное применение двух композиций:

$$R_q = I_q \circ S \circ I,$$
$$R_q = \sum_i \mu_{R_q}(d_k)/d_k.$$

Легко показать, что если на каждом этапе применять максиминную композицию [7], то релевантность документа I_d запросу I_q тоже будет вычисляться по формуле (2).

Таким образом, все три модели:

— модель текстового поиска, основанная на обобщении нечеткого отношения (п. 3.1),

— модель текстового поиска с расширением запроса, основанная на максиминной близости между нечеткими множествами (п. 3.2),

— модель текстового поиска из работы [7] дают одинаковые значения релевантности.

1. ПРИМЕР[3]

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*

Запрос включает в себя понятия c_1, c_2, c_3 и представлен вектором

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

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

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

индексирующее отношение I — матрицей:

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

[3] Из работы [7].

Тогда вектор релевантности по моделям из § 3 (формула (2)) имеет вид:

$$R_q = \begin{matrix} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10} \\ [0{,}5 & 1 & 1 & 0{,}3 & 0{,}4 & 0{,}5 & 1 & 0{,}9 & 0{,}7 & 0{,}5] \end{matrix}.$$

Вектор релевантности по модели из работы [6] имеет вид:

$$R_q = \begin{matrix} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10} \\ [0{,}4 & 0 & 0{,}7 & 0{,}2 & 0{,}06 & 0{,}27 & 0{,}7 & 0 & 0{,}27 & 0] \end{matrix}.$$

Нулевая релевантность запросу документов d_2, d_8, d_{10} объясняется отсутствием в этих документах понятий запроса, так как модель не предполагает расширение запроса.

Применяя α-срез при $\alpha = 0{,}5$, получаем

$$R_{q(.5)} = \begin{matrix} d_2 & d_3 & d_7 & d_8 & d_9 \\ [1 & 1 & 1 & 0{,}9 & 0{,}7] \end{matrix}$$

по моделям из § 3 и

$$R_{q(.5)} = \begin{matrix} d_3 & d_7 \\ [0{,}7 & 0{,}7] \end{matrix}$$

по модели из работы [6].

ЗАКЛЮЧЕНИЕ

Проанализированы модели текстового поиска в рамках теории нечетких множеств: две модели, предложенные в данной работе, и модель из работы [7]. Показано, что все три модели дают одну формулу вычисления релевантности документа запросу. Для сравнения моделей текстового поиска, основанных на теории нечетких множеств, с их четкими аналогами планируется экспериментальная проверка на реальных примерах.

ЛИТЕРАТУРА

1. *Онтологии* и тезаурусы / В.Д. Соловье в и др. — Казань, Москва, 2006. — 157 с.
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. — 384 с.
4. *Орловский С.А.* Проблемы принятия решений при нечеткой исходной информации. — М.: Наука, 1981. — 208 с.
5. *Mitrović Zoran*, *Rusov Srđan*. Z Similarity Measure Among Fuzzy Sets // FME Transactions. — 2006. — N 34. — P. 115—119.
6. *Knappe Rasmus.* Measures of Semantic Similarity and Relatedness for Use in Ontology-based Information Retrieval. — Denmark: Roskilde University, 2005. — 108 p.
7. *Karn Bhaskar.* Information retrieval system using fuzzy set theory — the basic concept. — URL: http://pchats.tripod.com/istebhaskar.pdf (дата обращения: 28.9.2012).

Статья представлена к публикации членом редколлегии О.П. Кузнецовым.

Людмила Александровна Панкова — канд. техн. наук, ст. науч. сотрудник, ☎ (495) 334-92-49, ✉ pankova@ipu.ru,

Валерия Александровна Пронина — канд. техн. наук, ст. науч. сотрудник, ☎ (495) 334-92-49, ✉ pankova@ipu.ru,

Институт проблем управления им. В.А. Трапезникова РАН, г. Москва.