OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

НЕЧЁТКИЕ СЕМАНТИЧЕСКИЕ МОДЕЛИ ТЕКСТОВОГО ПОИСКА

Панкова Л.А.*, Пронина В.А.**

Федеральное государственное бюджетное Учреждение науки Институт проблем управления им. В.А. Трапезникова Российской академии наук, г. Москва, Россия

***pankova@ipu.ru**

****pron@ipu.ru**

Понятия текстового поиска интерпретированы в терминах теории нечётких множеств. Предложены модели текстового поиска на основе теории нечётких множеств.
Ключевые слова: текстовый поиск, связанность, нечёткое отношение, релевантность, 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 рассматриваются 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 тексте документа веса определяются в автоматическом процессе концептуального индексирования [**Соловьев и др., 2006**]. Релевантность (семантическое соответствие) документа запросу в рассматриваемых моделях формально определяется с использованием отношения семантической связанности понятий. Семантическая связанность понятий может вычисляться формальным образом по онтологии предметной области данной коллекции или с использованием статистических методов, а может задаваться экспертом.

Текстовый поиск имеет дело с нечёткой априорной информацией, что не принимается в расчет в большинстве существующих чётких моделей (см., например, обзор из **[Панкова и др., 2011]**). Теория нечётких множеств даёт средства обращения с нечёткой информацией. В существующих работах по информационному поиску теория нечётких 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. Кроме того, необходимы дополнительные усилия, чтобы обнажить то, что теория нечетких множеств может предложить в этой области элегантные и интуитивно привлекательные методы.

В данной работе рассматриваются модели текстового поиска, использующие теорию нечётких множеств [**Заде, 2000**]. Предлагаемые модели используют принцип обобщения – универсальный принцип теории нечётких множеств [**Орловский, 1981**] – для перехода от отношений между понятиями к отношениям между документами и запросами, от связанности понятий к релевантности документов и запросов.

Структура работы. В первом разделе основные понятия текстового поиска интерпретируются в терминах теории нечётких множеств. Во втором разделе предложены модели текстового поиска в рамках теории нечётких множеств. В третьем разделе дан пример ранжирования коллекции документов по релевантности запросу, вычисленной предложенными методами.

1. Понятия текстового поиска в терминах теории нечётких множеств

Интерпретируем понятия текстового поиска в терминах теории нечётких множеств.

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 нечёткое 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.

1.2. Множество концептуальных индексов запросов можно представить как нечёткое 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:

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

1.3. Отношение семантической связанности понятий S можно представить как нечёткое 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. НЕЧЁТКИЕ МОДЕЛИ ТЕКСТОВОГО ПОИСКА, ОСНОВАННЫЕ НА ПРИНЦИПЕ ОБОБЩЕНИЯ

Предлагаемые модели используют принцип обобщения и интуитивно просты. Принцип обобщения – это универсальный принцип теории нечётких 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 нечёткое 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 нечёткое отношение связанности S' нечетких запросов с понятием обобщается, чтобы получить обобщённое нечёткое отношение связанности S'' на $\{I_q\} \times \{I_d\}$ с функцией принадлежности $\mu_{S''}(I_q, I_d)$, которую будем называть **обобщённой связанностью документов и запросов**:

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

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

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, обобщённое нечёткое отношение связанности документов с запросами S'' имеет вид:

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

Рисунки 1-3 иллюстрируют применение принципа обобщения.

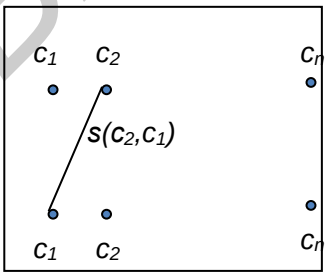

Рисунок 1 – Семантическая связанность понятий

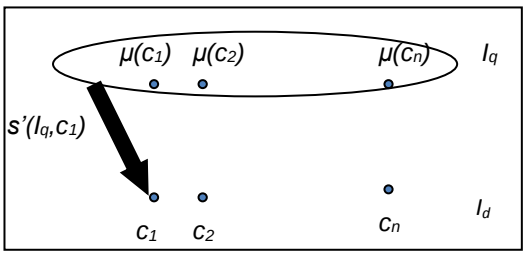

Рисунок 2 – Семантическая связанность запроса и понятия

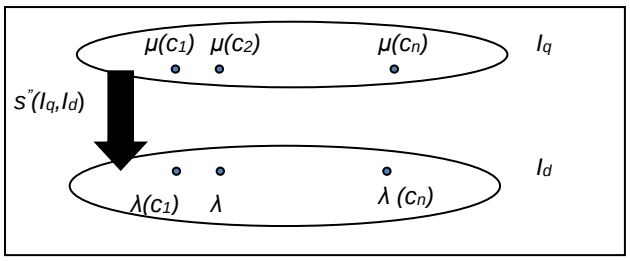

Рисунок 3 – Семантическая связанность запроса и документа

Если I_q и I_d – одноэлементные нечеткие множества, то S'' интерпретируется как обобщённое нечёткое отношение связанности двух нечётких понятий с функцией принадлежности (максимум исчезает, т.к. рассматривается одна пара понятий), которое будем называть **обобщённой 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 обобщённая связанность I_q и I_d можно представить как максимальное значение обобщённых связанностей всех пар нечётких понятий запроса и документа:

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

Вводится параметр **ширины связанности запроса и документа *N (q,d)***, определяющий число пар понятий запроса и документа, для которых обобщенная связанность принадлежит заданному полуинтервалу:

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

$$\tilde{c}_i\subset I_q,\tilde{c}_j\subset I_d,\ \ \delta_1<\delta_2,$$

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

Введённые понятия можно использовать для определения релевантности (семантического соответствия) документа запросу и ранжирования документов.

1. Релевантность документа запросу $R(q,d)$ вычисляется как обобщённая связанность документа и запроса, и документы ранжируются по значению релевантности:

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

2. Релевантность документа запросу вычисляется как ширина связанности $N(q,d)$ при заданном δ_1, и документы ранжируются по значению релевантности.

3. Для ранжирования документов используются $R(q,d)$ и $N(q,d)$ по алгоритму, описанному ниже.

- Для каждого документа вычисляется *R(q,d)*.
- Определяется α-срез – множество документов D_α, релевантных запросу с релевантностью большей α.
- На полуинтервале $(\alpha,\ \max_{d\in D}\{R(q,d)]$ вводится лингвистическая переменная «релевантность». Множество документов D_α разбивается на классы эквивалентности D_α^i в соответствии со значениями «релевантности»:

$$D_\alpha^i=\{d\,|\,\delta_{i1}<R(q,d)\le\delta_{i2},i=1,...,k\},$$

где k – число значений лингвистической переменной «релевантность».

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

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

- 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*

Запрос включает понятия c_1, c_2, c_3 и представлен вектором:

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

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

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

индексирующее отношение I – матрицей:

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

3.1. Ранжирование по обобщённой связанности документа и запроса

Релевантности документов запросу вычисляются как обобщённые связанности документов и запроса:

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

Определяется α-срез при α = .5:

$$R_{q(.5)}=\begin{bmatrix}d_2 & d_3 & d_7 & d_8 & d_9\\ 1 & 1 & 1 & .9 & .7\end{bmatrix}.$$

Документы упорядочиваются по значению $R_{q(.5)}$:

$$\begin{matrix}1\\.9\\.7\end{matrix}\begin{bmatrix}d_2 & d_3 & d_7\\ & d_8 & \\ & d_9 & \end{bmatrix}.$$

3.2. Ранжирование по ширине связанности запроса и документа

Релевантность документа запросу вычисляется как значение $N(q,d)$ при δ_1=.5:

$$N(q,d)=\begin{bmatrix}d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10}\\ 0 & 2 & 2 & 0 & 0 & 0 & 3 & 1 & 1 & 0\end{bmatrix}.$$

Документы упорядочиваются по значению $N(q,d)$:

$$\begin{matrix}3\\2\\1\end{matrix}\begin{bmatrix}d_7 & \\ d_2 & d_3\\ d_8 & d_9\end{bmatrix}.$$

3.3. Ранжирование по обобщённой связанности **и** по **ширине связанности**

Вычисляется α-срез при α = .5 (по 3.1.).

$$R_{q(.5)}=\begin{bmatrix}d_2 & d_3 & d_7 & d_8 & d_9\\ 1 & 1 & 1 & .9 & .7\end{bmatrix}.$$

Полуинтервал (.5, 1] делится на два полуинтервала (.5, .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 $N^i(q,d)$. Итоговая ранжировка имеет вид:

$$\begin{matrix}(.9,1]\\(.9,1]\\(.9,1]\\(.5,.9]\end{matrix}\begin{matrix}3\\2\\1\\1\end{matrix}\begin{bmatrix}d_7 & \\ d_2 & \\ d_3 & \\ d_8 & d_9\end{bmatrix}.$$

Заключение

Понятия текстового поиска интерпретируются в терминах теории нечётких множеств. Предлагаются модели текстового поиска в рамках теории нечётких множеств. Дан пример вычисления релевантности коллекции документов запросу по предложенным моделям.

Для проверки адекватности предложенных моделей текстового поиска планируется экспериментальное исследование на коллекции научно-технических текстов.

Библиографический список

[**Соловьев и др., 2006]** 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.

[Панкова и др., 2011] 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.

[Заде, 2000] Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. СПб.: Питер, 2000.

[**Орловский**, **1981**] Орловский С.А. Проблемы принятия решений при нечеткой исходной информации. М: Наука. Главная редакция физико-математической литературы, 1981.

SEMANTIC TEXT RETRIEVAL BASED ON FUZZY SET THEORY

Pankova L.A. *, Pronina V.A. **

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Profsoyuznaya ul., 65, Moscow, 117997 Russia

***pankova@ipu.ru**

****pron@ipu.ru**

Text retrieval concepts are interpreted in terms of the fuzzy set theory. The text retrieval models based on the fuzzy set theory are proposed.

Introduction

The crisp models of text retrieval does not count fuzziness of information. The fuzzy set theory provides means of handling fuzzy information. In the existing works the fuzzy set theory is mainly used to represent the ontology and the implementation of more flexible ways for formulating queries. In practice, text retrieval models are still based on other approaches. The fuzzy set theory can offer in this area are elegant and intuitively attractive methods.

Main Part

The proposed models are based on the generalization principle and intuitively simple. The generalization principle is universal principle of the fuzzy set theory.

In proposed models the generalization principle is used for moving from the relation on the concepts to the relation on documents and queries, namely, from relatedness of concepts to the relevance of documents and queries.

The article presents the example that shows results of modeling.

Conclusion

There is planned experimental verification on a collection of scientific-technical texts.