УДК 004.048+025.4.03

А.Н. Целых, Э.М. Котов, А.А. Целых

МЕТОД ИНФОРМАЦИОННОГО ПОИСКА НА ОСНОВЕ НЕЧЕТКОГО СХОДСТВА СИТУАЦИЙ

Рассматривается подход к организации нечеткого информационного поиска в ситуации, когда запрос выражен нечетким множеством, определенным на множестве поисковых индексов. Для установления отношений между документом и запросом используется нечеткий тезаурус, что позволяет для некоторого запроса идентифицировать релевантные документы, которые иначе не были бы выданы. Для семантической сети, включающей в себя ситуации, концепты и поисковый запрос, рассматривается процедура расширения и определения релевантности ситуации запросу. Связывая вершину-запрос с определяющими ее концептами, получим теоретико-графовую структуру, на основе которой удобно определять релевантность запроса каждому из объектов. Рассматривается метод, основанный на определении степени нечеткого сходства запроса (поискового образа) с ситуацией. Релевантность вычисляется методом минимального значения. Для расчета степени нечеткого включения запроса в каждую из ситуаций используется формула нечеткого сходства по Заде. Ситуация, для которой степень включения запроса наибольшая, выбирается в качестве искомого решения.

Неопределенность; неточность; нечеткость; информационный поиск; релевантность.

A.N. Tselykh, E.M. Kotov, A.A. Tselykh

METHOD OF INFORMATION RETRIEVAL BASED ON A FUZZY SIMILARITY OF SITUATIONS

This paper considers an approach to the problem of fuzzy information retrieval in a situation where a query is given by a fuzzy set defined on a set of search indexes. To find the relationship between a document and a query, we use a fuzzy thesaurus that allows for some query to identify relevant documents that otherwise would not be issued. For a semantic network that includes situations, concepts and a search query, we consider a procedure for extension and determining the relevance of a situation to a request. By linking the query vertex with defining concepts, we obtain a graph theoretical structure that is used to determine the relevance of the query to each object. We also consider the method based on determining the degree of similarity of a fuzzy query (search image) and the situation. We compute the relevance score with a method of a minimum value. To calculate the degree of fuzzy inclusion of a query in each of the situations, we use the formula of fuzzy similarity by Zadeh. The situation with the highest degree of inclusion is selected as a desired solution.

Vagueness; imprecision; fuzziness; information; retrieval.

Быстрое развитие средств сбора, хранения и распространение информации делает разработку систем, которые управляют информационными потоками и извлекают информацию, соответствующую потребностям пользователя, важной проблемой.

Первичная цель любой информационно-поисковой системы (рис. 1) состоит в помощи пользователям эффективно получить желаемую информацию. Большинство коммерческих информационно-поисковых систем в настоящее время все еще строится на основе поисковой модели, использующей булеву логику. Однако эти системы обладают определенными ограничениями, так как не способны представить нечеткую, неопределенную информацию. Если в запросе присутствует нечеткая информация, то обработка запроса такими системами не осуществляется должным образом [1].

Нечеткие информационно-поисковые системы базируются на технологиях, использующих нечеткую логику и нечеткие отношения с целью получения наилучшего результата, соответствующего пользовательскому запросу. В отличие от булевых систем, нечеткие системы оперируют с данными, которые отражают степень (меру) принадлежности некоторого элемента *x* нечеткому множеству *A*.

Понятие нечеткого отношения возможно считать одним из основных понятий теории нечетких множеств. Эти отношения позволяют формализовать неточные утверждения «*x* значительно больше чем *y*» или «*x* почти равно *y*». По сравнению с вероятностным, нечеткий метод, при использовании в информационно-поисковых системах, позволяет резко сократить объем производимых вычислений, что, в свою очередь, приводит к увеличению быстродействия нечетких поисковых систем [2].

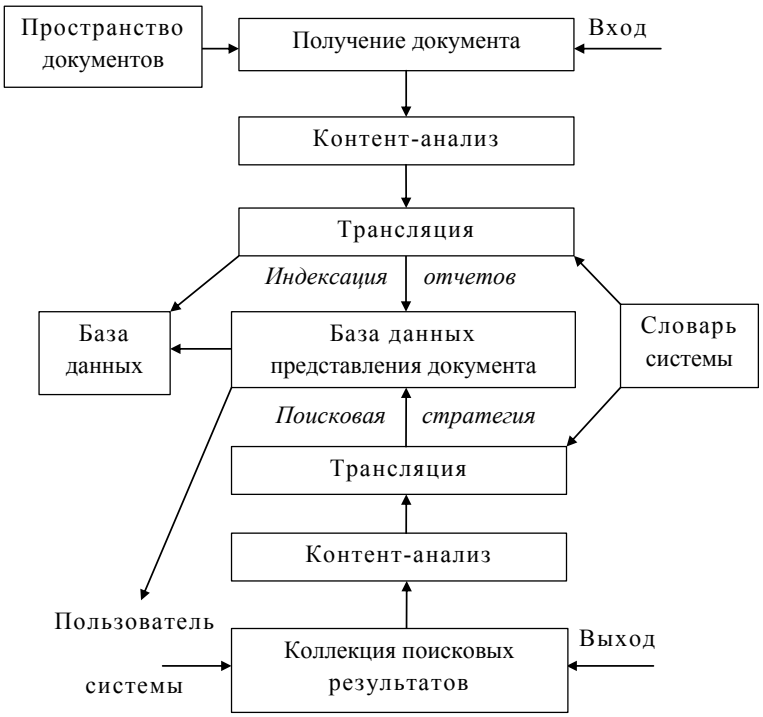

Рис. 1. Функции ИПС

В общем случае можно говорить, что информационно-поисковая система состоит из двух частей: текстовый архив данных, который является множеством текстовых единиц (документов), и непосредственно поисковые средства. Пользователь представляет поисковой системе запросы, описывающие потребность в тех или иных видах документов. Поисковая система определяет соответствие запросов в индексной базе документов в текстовом архиве данных. В результате пользователю возвращается ранжированный список коллекции документов, который поисковая система считает «наилучшим результатом».

Таким образом, мы можем определить информационный поиск как проблему выбора документальной информации из источника хранения в ответ на поисковый запрос, т.е. установление соответствия слов или других символов запроса тем, которые характеризуют отдельный документ.

Неопределенность может присутствовать при различных ситуациях: неопределенность соотношения «известного»/«неизвестного» в предмете поиска; неопределенность системы характеристических признаков для структуризации предмета поиска; лексическая неопределенность как фактор степени соответствия информационно-поискового языка естественнонаучному языку предметной области; неопределенность критериев сравнения; неопределенность интерпретации поисковых образов документов; неопределенность тематического соответствия и степени полноты представления проблематики.

Модель информационного поиска включает в себя два конечных множества: множество поисковых индексов, содержащих информацию о документах $X = \{x_1, x_2, \dots, x_n\}$, и множество релевантных документов $Y = \{y_1, y_2, \dots, y_n\}$.

При нечетком информационном поиске релевантность поисковых индексов к отдельным документам выражена нечетким отношением

$$R = X \times Y \to [0,1],$$

таким образом, значение $R(x_i, y_j)$ определяет для каждого $x_i \in X$ и $y_j \in Y$ степень релевантности поискового индекса x_i документу y_j.

Нечеткий тезаурус играет важную роль в установлении отношений между документом и запросом. Нечеткий тезаурус – рефлексивное отношение *T*, определенное на множестве *X*. Для каждой пары поискового индекса $\langle x_i, x_k \rangle \in X$, T выражает ассоциацию x_i с x_k – это степень, с которой значение поискового индекса X_k совместимо со значением данного поискового индекса x_i. Подобное отношение должно сталкиваться с проблемой наличия синонимов среди множества поисковых индексов. В результате отношения помогают идентифицировать релевантные документы для некоторого запроса, которые иначе не были бы идентифицированы. Это происходит всякий раз, когда документ характеризуется поисковым индексом, который синонимичен с поисковым индексом, содержавшимся в запросе.

При нечетком информационном поиске запрос может быть выражен нечетким множеством, определенным на множестве поисковых индексов *X*. Обозначим через *A* некоторое нечеткое множество, представляющее отдельный запрос. Тогда, при сравнении с нечетким тезаурусом *T*, мы получаем новое нечеткое множество (обозначим его через *B*), которое представляет собой дополненный запрос (т.е. множество, дополненное связями с поисковыми индексами).

Пусть задана семантическая сеть, которая включает в себя три ситуации [3] – y_1, y_2, y_3, шесть концептов – $x_1, x_2, \dots, x_6$ и запрос z, относящийся к трем концептам. Поиск начинается с одного из исходных узлов (концептов), которому соответствует запрос, например, с x_3 (рис. 2). Проследим процедуру расширения и определим релевантность ситуации y_2 запросу, а именно $v(RZ(y_2, z))$. Обозначим связи дугами, которые пометим весами, показывающими силу семантической связи между соединенными узлами.

Связывая вершину-запрос z с определяющими ее концептами $x \in X$, получаем структуру, с помощью которой удобно определять релевантность запроса z каждому объекту из *Y*.

Представляется интересным исследовать подход, основанный на определении степени нечеткого сходства запроса (поискового образа) z и ситуации y.

Соответствие каждого описания ситуации $y \in Y$ при помощи $x \in$ X определяется как нечеткое подмножество множества X концептов. Аналогично, каждое описание запроса z может определяться как такое же нечеткое подмножество. Более того, если концепт есть некоторая лингвистическая переменная, то ребро

(x_i, y) может быть взвешено уже не числом от 0 до 1, а некоторым значением лингвистической переменной x, т.е. нечетким множеством. Каждая ситуация y может характеризоваться нечетким множеством второго уровня [4]. То же самое справедливо и для запроса z.

При использовании вершины-запроса z получим следующий граф соответствий (см. рис. 2).

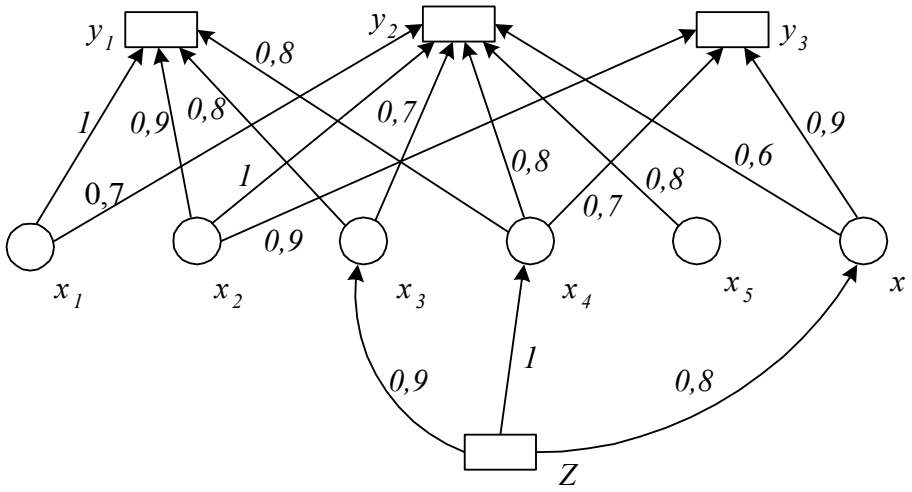

Рис. 2. Граф соответствий с присоединенным запросом

Для формализации процедуры вычисления релевантности $d(y_j, z)$, степени нечеткого включения $z \to y_j$ и степени нечеткого сходства $z \leftrightarrow y_j$ представим описание каждой ситуации $y_j, y_j \in Y$ и запроса z в виде векторов. Имеем

$$y_1 : A_1 = (1, 0.9, 0.8, 0.8, 0, 0);$$
$$y_2 : A_2 = (0.7, 1, 0.7, 0.8, 0.8, 0.6);$$
$$y_3 : A_3 = (0, 0.9, 0, 0.7, 0, 0.9);$$
$$z : B = (0, 0, 0.9, 1, 0, 0.8).$$

Отсюда, в соответствии с процедурой вычисления значения релевантности $d(y, z)$ методом минимального значения, получим

$$d(y_1, z) = \frac{1}{N} \sum_i \min(a_i, b_i) = \frac{1}{3}(0,8 + 0,8 + 0) = \frac{1,6}{3} = 0,53,$$

где N – количество ненулевых элементов вектора описания z.

$$d(y_2,\ z) = \frac{1}{N} \sum min\ (a_i,\ b_i) = \frac{1}{3}(0,7 + 0,8 + 0,6) = 0,7;$$

$$d(y_3,\ z) = \frac{1}{3}(0 + 0,7 + 0,8+) \approx 0,5.$$

Определим теперь степень нечеткого включения запроса z в каждую из ситуаций $y \in Y$. Для этого воспользуемся формулой

$$v(z \to y) = v(B \to A) = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} (b_i \to a_i),$$

где $b \to a = \max(1 - b, a)$ (по логике Заде) [5].

Для заданных векторов A_1, A_2, A_3, B получим:

$$\nu_1 (B \to A_1) = \frac{1}{6}\sum(1+1+0,8+0,8+1+0,2) = \frac{1}{6}\cdot 4,8 = 0,8;$$

$$\nu_2 (B \to A_2) = \frac{1}{6}\sum(1+1+0,7+0,8+1+0,6) = \frac{1}{6}\cdot 5,1 = 0,85;$$

$$\nu_3 (B \to A_3) = \frac{1}{6}\sum(1+1+0,8+0,8+1+0,2) = \frac{1}{6}\cdot 4,8 = 0,8.$$

Та ситуация *y*, для которой степень ν включения запроса *z* наибольшая, может быть выбрана в качестве искомого решения.

Ранее в работе [6] предложена модель для поиска оптимальных решений с использованием нечеткой семантической сети, основанная на определении степени принадлежности текущей ситуации к тому или иному классу эталонных ситуаций, использующих понятие степени нечеткого сходства. В данном исследовании предложен подход к организации нечеткого информационного поиска в ситуации, когда запрос выражен нечетким множеством, определенным на множестве поисковых индексов. При расчете степени нечеткого включения запроса в каждой из ситуаций используется формула нечеткого сходства по Заде. В дальнейшем представляется интересным вычислить оценку релевантности ситуации поисковому образу в семантической сети с присоединенной вершиной.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Bordogna G., and Pasi G.* Handling Vagueness in Information Retrieval Systems // Proceedings of the Second New Zealand International Two-Stream Conference on Neural Networks and Expert Systems, Dunedin, Nuova Zelanda, 20-23 November 1995. – P. 110-116.
2. *Liu Z.* Information Retrieval Using Relevance Feedback for the Mobile Internet. Thesis, University of North Dakota, May 2006.
3. *Мелихов А.Н., Берштейн Л.С., Коровин С.Я.* Ситуационные советующие системы с нечеткой логикой. – М.: Наука, 1990. – 272 с.
4. *Ларичев О.И.* Анализ процессов принятия человеком решений при альтернативах, имеющих оценки по многим критериям // Автоматика и телемеханика. – 1981. – № 8. – С. 131-141.
5. *Заде Л.А.* Понятие лингвистической переменной и его применение к принятию приближенных решений. – М.: Мир, 1976. – 168 с.
6. *Целых А.Н., Котов Э.М.* Методы нечетко-множественного анализа и моделирования социальных графов // Современные проблемы науки и образования. – 2013. – № 6. URL: www.science-education.ru/113-11178.

Статью рекомендовал к опубликованию д.т.н., профессор В.П. Карелин.

Целых Александр Николаевич – Южный федеральный университет; e-mail: ant@sfedu.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: +79185562047; кафедра информационно-аналитических систем безопасности; д.т.н.; профессор.

Котов Эдуард Михайлович – e-mail: emkotov@sfedu.ru; тел.: +79885887317; кафедра информационно-аналитических систем безопасности; ассистент.

Целых Алексей Александрович – e-mail: tselykh@sfedu.ru; тел.: +79185116226; кафедра информационно-аналитических систем безопасности; к.т.н.; доцент.

Tselykh Alexander Nikolaevich – Southern Federal University; e-mail: ant@sfedu.ru; 44, Nekrasovsky, Taganrog, 347928, Russia; phone: +79185562047; the department of information and analytical systems security; dr. of eng. sc.; professor.

Kotov Eduard Michaylovich – e-mail: emkotov@sfedu.ru; phone: +79885887317; the department of information and analytical systems security; assistant.

Tselykh Alexey Alexandrovich – e-mail: tselykh@sfedu.ru; phone: +79185116226; the department of information and analytical systems security; cand. of eng. sc.; assistant professor.