МОДЕЛИРОВАНИЕ ИНФОРМАЦИОННОГО ПОИСКА В ИНТЕРНЕТ НА БАЗЕ НЕЧЕТКИХ ЗНАНИЙ

В.А.КАСУМОВ

Информационно-поисковые системы Интернета, содержащие тематические каталоги, дают более точные результаты. Несмотря на это, они имеют немало недостатков, так, например, создание и ведение тематического каталога выполняется человеком вручную, а это требует огромную интеллектуальную работу и профессионализм персонала, обновление базы данных такой системы является очень трудоемким и длительным процессом, с увеличением количества подкаталогов в тематическом каталоге усложняет процесс поиска нужных источников и достижения желаемых результатов. Поисковые системы, основанные на автоматические индексы, наоборот являются более гибкими и простыми для реализации, охватывают больше информационных источников, чем тематические каталоги, т.е. полнота автоматических индексов намного превышает полноту тематических каталогов. Однако точность результатов поиска при этом относительно отстает.

Исследования показывают, что разработка методов создания автоматических тематических каталогов без вмешательства человека является актуальной и много обещающей. Здесь предполагается, что поисковый робот системы сам должен определять тематику информационного ресурса, наиболее подходящие тематические профили из тематического каталога и включить ресурс в них. При этом созданный тематический каталог может быть построен согласно существующим стандартам УДК, ББК и т.д. Созданный таким образом автоматический тематический каталог может заменить человека (администратора или оператора поисковой системы), позволит выполнить работу по созданию, обслуживанию и обновлению тематического каталога поисковой системы.

В настоящей статье исследуется модель информационного поиска. Рассматриваются методы определения тематики информационных ресурсов, разбиения информационного массива по тематическим направлениям и улучшения тематического каталога. Для решения поставленных задач используется теория нечетких множеств и отношений, а также подход Беллмана-Заде.

1. Введение

Для построения эффективных информационно-поисковых систем, необходимо изучить информационную среду поиска в целом, разработать методов тематического разбиения, индексирования, поиска и представления информационных ресурсов и запросов. Модель поисковой системы должна включить в себя как необходимые множества, так и отношения между этими множествами [1-2].

Поисковые системы, ориентированные на тематические каталоги имеют более высокие показатели точности, чем ориентированные на автоматические индексы. Однако полнота автоматических индексов обычно намного превышает полноту тематических каталогов. Первые более гибкие и легко подаются адаптацию, т.е. программы-индексаторы (роботы, пауки, спайдеры и др.) периодически без особого труда могут обновлять базу [3].

Обычно методы индексирования используют два механизма: приписка терминов из тематического рубрикатора по смысловому анализу и извлечение терминов (соответствующих тематическому рубрикатору) из тела

информационных ресурсов. Первый подход требует применение мощного семантического аппарата или ручное индексирование специалистами в области тематики индексируемого ресурса. Методы, основанные на семантический анализ, трудно реализовать и, вообще, мало развиты. Ручное индексирование является более точным, однако, создание хорошего тематического каталога зависит от профессионализма персонала и требует огромную интеллектуальную работу [4,5]. Традиционно, работу по созданию и адаптации тематических каталогов выполняет администратор или оператор поисковой системы ручным способом. Далее предполагается, что индексирование проводится с помощью методов второго типа, т.е. термины извлекаются из содержания информационных ресурсов согласно тематическому рубрикатору.

Существуют достаточно хорошие методы, позволяющие извлечь из содержания информационного ресурса наиболее важные термины и вычислить коэффициентов их важности [4,6]. Примером таких методов является статистический метод, с помощью которого достигается хороший результат, если при этом использовать справочник исключений «стоп-слов», т.е. справочника служебных слов, глаголов, местоимений и т.д. [7].

Слабым местом такого подхода является следующее: разные авторы могут использовать разные термины, смысл которых очень близки (возможно идентичны), например, “information retrieval” и “information search”; в индексируемом источнике выделяется важное ключевое слово, которое является не распространенным термином в данной области или не входит тематический рубрикатор, а синонимы данного термина могут оказаться значимыми терминами; в информационном ресурсе вместо термина используется английские, латинские или другие эквиваленты; один термин может ассоциировать достаточно близкие по тематике другие термины; языки источников по одной тематике могут быть разными [1,8,9].

Без учета выше перечисленных обстоятельств нельзя достичь достаточного уровня индексирования, от которого непосредственно зависит результат поиска. Выход из положения является создание и использование словарей синонимов и ассоциирующих слов, а также переводчиков [1,9].

Поэтому исследование проблемы создания автоматических тематических каталогов является много обещающим. Для решения данной проблемы требуется автоматизировать работу распределения документов по тематическим каталогам путем определения тематики документа и нахождения тематических каталогов, имеющие наиболее релевантный профиль [1].

В настоящей статье излагаются методы, позволяющие создать автоматические тематические каталоги без вмешательства человека и улучшения качества тематических каталогов с применением синонимов и ассоциирующих терминов, а также поиска наиболее релевантных информационных ресурсов и тематических каталогов к запросам пользователей. С этой целью используется аппарат нечетких множеств и подход Беллмана-Заде [10].

2. Нечеткая модель информационного пространства Интернет

Пусть $D=\{d_i\}_I$ – множество информационных ресурсов Интернет, которое требуется разбить на подмножества (тематические каталоги) по тематике, аналогично библиотечно-информационной системе. Множество классификационных направлений, которых назовем тематическими каталогами, определяемыми своими тематическими профилями, обозначим через $K=\{K_l\}_L$. Каждый тематический профиль определяется собственными своими дескрипторами, ключевыми словами или другими лексическими единицами, которых назовем терминами. Множество терминов, определяющие тематические направления тематического профиля K_l, обозначим через T_l^K. Необходимо отметить, что эти множества могут частично пересекаться, т.е.:

$$T_i^K \cap T_j^K \neq \varnothing, \text{ для } \forall\ i \neq j. \qquad (1)$$

Объединение множеств терминов всех тематических профилей поисковой системы составляет множество терминов системы. Если через $T=\{t_j\}_J$ обозначим множество терминов поисковой системы, тогда

$$T = \bigcup_{l=1}^{L} T_l^K . \qquad (2)$$

В качестве множества терминов можно использовать любой универсальный библиографический классификатор, такие как УДК, ББК и т.д. Возможен другой подход, который лежит на основе многих современных информационно-поисковых систем Интернета, суть которого заключается в создании множества терминов самой информационно-поисковой системой или ее администратором, как база индексов или метаданных, и дополнении в процессе функционирования системы.

Введем еще одно множество - множество синонимов и ассоциирующих слов терминов $t_j \in T, j=\overline{1,J}$, которое представляется в виде нечеткого отношения. Для простоты будем предполагать, что синонимы и ассоциирующие термины составляют одно множество $S=\{s_v\}_V$.

Учитывая выше сказанное информационный поиск можно представить в виде пятерки $IR=\{K,D,T,Q,R\}$, где IR является результатом информационного поиска (Information Retrieval) и выдается пользователю в виде вектора (списка) названий, адресов и других реквизитов источников информации, Q - запросы пользователей, которые рассматриваются ниже. $R=\{R^D, R^K, R^S, R^Q\}$ - множество отношений, которое определяет отношения типа “информационный ресурс – термин” (R^D), “тематический каталог – термин” (R^K), “термин - синоним” (R^S) и “ запрос – термин” (R^Q). Необходимо отметить, что все отношения, определяемые здесь и рассматриваемые далее, являются нечеткими.

На рис.1 информационный поиск представлен в виде направленного графа, вершинами которого являются множества, а дугами представляются отношения между ними. В данный граф включены все множества и отношения информационного поиска.

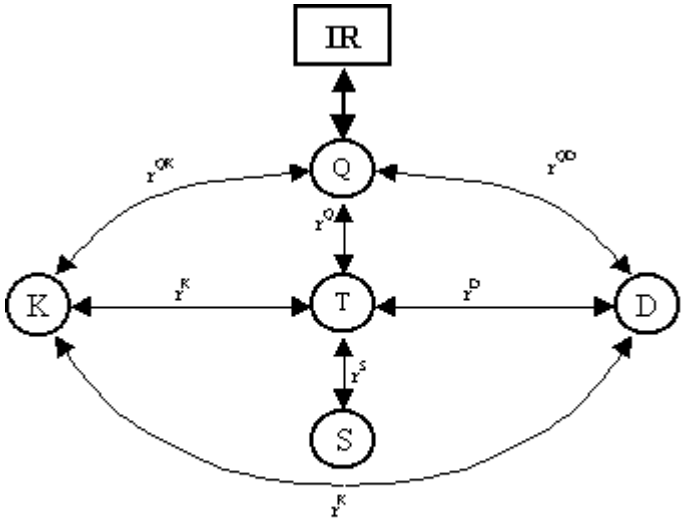

Рис.1. Представление информационного поиска в виде ориентированного графа

Отношение между множествами тематических профилей и терминов называется тематическим рубрикатором. Как было сказано выше, каждый тематический профиль определяется множеством терминов, где одни термины могут быть релевантными одному тематическому профилю, другие термины - нескольким тематическим профилям. Тематический рубрикатор представляется в виде нечеткой реляционной матрицы размерности LxJ, строки которой соответствуют тематическим профилям, а столбцы - терминам. Таким образом, каждый тематический рубрикатор задается в следующем виде:

$$K_l = \{t_j / \varphi_{lj}\},\ l = \overline{1,L},\ j = \overline{1,J},\ , \qquad (3)$$

где φ_{lj} - функция принадлежности термина к тематическому профилю, т.е. степени релевантности термина t_j тематическому каталогу K_l. Тогда R^K - отношение между множеством тематических профилей и множеством терминов имеет вид:

$$r_l^K = \left\{\varphi_{lj} : T \times K \rightarrow [0,1]\right\},\ l = \overline{1,L},\ j = \overline{1,J},\ R^K = \left\{r_l^K\right\}_L . \qquad (4)$$

Значения элементов матриц отношений R^S (рассматривается далее) и R^K определяются в начальном этапе создания информационно-поисковой системы, для чего применяется метод экспертных оценок. В процессе функционирования системы значения отношений φ_{lj} подвергаются адаптации, т.е. обучению.

Теперь рассмотрим, как представляются в поисковой системе проиндексированные информационные ресурсы. Как выше отметили, в результате индексирования каждому ресурсу информационного пространства привязываются некоторые термины из множества T. Терминам привязываются весовые коэффициенты относительно каждого информационного ресурса, которые определяют степень важности термина для информационного ресурса.

Множество терминов (ключевых слов) информационного ресурса d_i обозначим через T_i^D, которое является нечетким множеством. Весовые коэффициенты терминов относительно каждого информационного ресурса определяются функцией принадлежности термина в множество T_i^D, которая получает значения в пределе [0,1]. Все множества T_i^D нормируются, т.е. дополняются терминами множества T, отсутствующими в множестве T_i^D, функция принадлежности которых в T_i^D получает нулевые значения.

Таким образом, множество информационных ресурсов D можно представить в виде нечеткой реляционной матрицы размерности IxJ. Элемент,

находящиеся в пересечении i-ой строки и j-го столбца определяет вес термина t_j для информационного ресурса d_i, т.е.:

$$d_i = \{t_j / \mu_{ij}\},\ i = \overline{1,I},\ j = \overline{1,J}\ , \qquad (5)$$

а отношение между элементами d_i и t_j определяется следующим образом:

$$R_i^D = \left\{\mu_{ij} : D \times T \rightarrow [0,1]\right\},\ i = \overline{1,I},\ j = \overline{1,J}\ ,\ R^D = \left\{R_i^D\right\}_I , \qquad (6)$$

где μ_{ij} - функция принадлежности термина t_j в множество терминов T_i^D, значения которой определяются в результате индексирования информационного ресурса.

3. Разбиение информационного пространства на тематические каталоги по профилям ресурсов

Теперь рассмотрим задачу распределения информационных ресурсов Интернет по тематическим направлениям, т.е. разбиения информационных ресурсов на тематические каталоги. Если T_i^D - множество терминов информационного ресурса d_i и $T = \bigcup_{l=1}^{L} T_l^K$ - тематический рубрикатор, тогда релевантность определяется в виде пересечения этих множеств [10]:

$$T \cap T_i^D = \left(\bigcup_{l=1}^{L} T_l^K\right) \cap T_i^D = \bigcup_{l=1}^{L} \left(T_l^K \cap T_i^D\right). \qquad (7)$$

Для выявления наиболее релевантного тематического каталога для ресурса d_i можно использовать подход Беллмана-Заде.

На первом этапе определяется степени релевантности информационного ресурса d_i каждому тематическому каталогу множества K по отношению всех терминов данного ресурса. Как отметили выше, ресурс d_i в системе представляется отношением R_i^D данного ресурса к терминам множества T:

$$R_i^D = \{\mu_{i1}, \mu_{i2}, ..., \mu_{iJ}\},\ i = \overline{1,I}\ . \qquad (8)$$

Аналогично представляются тематические каталоги:

$$R_l^K = \{\varphi_{l1}, \varphi_{l2}, ..., \varphi_{lJ}\},\ l = \overline{1,L}\ . \qquad (9)$$

Тогда релевантность информационного ресурса d_i тематическому каталогу K_l можно рассмотреть как пересечение множеств отношений R_i^D и R_l^K, т.е. $R_{il}^{DK} = R_i^D \cap R_l^K$. Как известно, пересечение нечетких множеств определяется как алгебраическое произведение соответствующих элементов этих множеств [10], т.е. если $\eta_{ij}^l \in R_{il}^{DK}$, $l = \overline{1,L}$, тогда

$$\eta_{ij}^l = \mu_{ij} \cdot \varphi_{lj},\ l = \overline{1,L},\ j = \overline{1,J}\ . \qquad (10)$$

Здесь η_{ij}^l - степень релевантности тематики информационного ресурса d_i к профилю тематического каталога K_l по отношению термина t_j.

Далее определим наиболее предпочтительный (недоминируемый) тематический каталог. Пусть K^{ab} - абстрактный тематический каталог, объединяющий в себя все наилучшие отношения релевантности информационного ресурса d_i ко всем профилям тематического каталога по всем его терминам. K^{ab} можно определять путем объединения всех нечетких

множеств R_{il}^{DK} для всех $l=\overline{1,L}$, т.е. нахождения максимумов среди η_{ij}^{l} по всем терминам для всех каталогов [10]:

$$\eta_{ij}^{ab}=\max_{l=1,L}\{\eta_{ij}^{l}\},\ j=\overline{1,J}, \qquad (11)$$

где η_{ij}^{ab} - степень релевантности d_i к K^{ab} по отношению термина t_j. Исходя из этого, можно сказать, что K^{ab} представляет собой абстрактный каталог, являющийся наиболее релевантным ресурсу d_i.

Теперь найдем каталог K_l^* из множества $\{K_l\}_L$, тематический профиль которого является наиболее близким профилю абстрактного каталога K^{ab} относительно всех терминов t_j информационного ресурса d_i. С этой целью вычислим суммарное среднеквадратическое отклонение коэффициентов релевантности всех тематических каталогов K_l от коэффициентов релевантности K^{ab} по следующей формуле:

$$\lambda_{il}=\frac{\bar{J}}{J}\sum_{j=1}^{J}\left(\eta_{ij}^{ab}-\eta_{ij}^{l}\right)^2,\ l=\overline{1,L}, \qquad (12)$$

где $\bar{J}$-количество терминов информационного ресурса d_i, весовые коэффициенты которых поучают не нулевые значения для тематического каталога K_l.

Отсюда видно, что тематический каталог, имеющий минимальное среднеквадратическое отклонение является наиболее предпочтительным тематическим каталогом для данного информационного ресурса:

$$\lambda_i^*=\min_{l=1,L}\{\lambda_{il}\}. \qquad (13)$$

Таким образом, тематический каталог K_l^*, имеющий минимальное отклонение λ_i^* является наилучшим из множества $\{K_l\}_L$, т.е. профиль K_l^* наиболее близко профилю K^{ab} и, соответственно, наиболее подходящий тематике информационного ресурса. Отсюда следует, что данный ресурс необходимо включить в тематический каталог K_l^*.

Исходя из (1), можно предположить, что каждый информационный ресурс может оказаться релевантным не только к одному тематическому каталогу, а к нескольким. Для определения всех наиболее подходящих каталогов введем пороговые значения $\delta_l, l=\overline{1,L}$ для степени релевантности тематических каталогов. Таким образом, вытекает следующая задача: определять все наиболее предпочтительные тематические каталоги, степень релевантности которых не ниже пороговых значений.

Для определения наиболее предпочтительных каталогов необходимо из (12) найти все тематические каталоги, для которых отклонение тематики λ_{il} не превышает пороговое значение, т.е. удовлетворяют условие:

$$\lambda_{il}\le\delta_l,\ l=\overline{1,L}. \qquad (14)$$

Для наглядной иллюстрации предложенного метода и доказательства адекватности рассмотрим пример.

Пример 1. Пусть множество терминов состоит из десяти (J=10) терминов, а тематический рубрикатор включает в себя четыре каталоги (L=4), т.е.

$T=\{t_1,t_2,t_3,t_4,t_5,t_6,t_7,t_8,t_9,t_{10}\}$ и $K=\{K_1,K_2,K_3,K_4\}$. Отношение каталогов K_l к терминам t_j задано нечеткой реляционной таблицей R_1^K:

$$K_1(t_j/\varphi_{1j})=\{t_1/0.75;\ t_2/0;\ t_3/0.63;\ t_4/0;\ t_5/0.95;\ t_6/0.82;\ t_7/0;\ t_8/0;\ t_9/0.78;\ t_{10}/0\},$$

$$K_2(t_j/\varphi_{2j})=\{t_1/0;\ t_2/0.8;\ t_3/0.79;\ t_4/0.65;\ t_5/0.6;\ t_6/0;\ t_7/0.9;\ t_8/0;\ t_9/0.69;\ t_{10}/0.78\},$$

$$K_3(t_j/\varphi_{3j})=\{t_1/0.56;\ t_2/0.83;\ t_3/0;\ t_4/0;\ t_5/0.64;\ t_6/0;\ t_7/0.76;\ t_8/0.95;\ t_9/0;\ t_{10}/0.69\},$$

$$K_4(t_j/\varphi_{4j})=\{t_1/0;\ t_2/0.78;\ t_3/0.68;\ t_4/0.54;\ t_5/0.8;\ t_6/0.2;\ t_7/0.86;\ t_8/0;\ t_9/0.71;\ t_{10}/0.9\}.$$

Допустим, поисковая система проиндексировала информационный ресурс d_1 и выявила следующие коэффициенты важности терминов для данного ресурса:

$$d_1(t_j/\mu_{1j})=\{t_1/0;\ t_2/0.9;\ t_3/0.71;\ t_4/0.87;\ t_5/0.54;\ t_6/0;\ t_7/0.6;\ t_8/0;\ t_9/0.95;\ t_{10}/0.58\}.$$

Тематические каталоги и информационный ресурс можно представить в табличном виде:

$$R_l^K=\begin{pmatrix}0.75 & 0 & 0.63 & 0 & 0.95 & 0.82 & 0 & 0 & 0.78 & 0\\ 0 & 0.8 & 0.79 & 0.65 & 0.6 & 0 & 0.9 & 0 & 0.69 & 0.78\\ 0.56 & 0.83 & 0 & 0 & 0.64 & 0 & 0.76 & 0.95 & 0 & 0.69\\ 0 & 0.78 & 0.68 & 0.54 & 0.8 & 0.2 & 0.86 & 0 & 0.71 & 0.9\end{pmatrix}$$

и

$$R_1^D=(0\quad 0.9\quad 0.71\quad 0.87\quad 0.54\quad 0\quad 0.6\quad 0\quad 0.95\quad 0.58).$$

Вычислим значения η_{1j}^l по формуле (10), т.е. $\eta_{ij}^l=\mu_{ij}\cdot\varphi_{lj},\ l=\overline{1,4},\ j=\overline{1,10}$. Получим:

$$R_{jl}^{DK}=\begin{pmatrix}0 & 0 & 0.45 & 0 & 0.51 & 0 & 0 & 0 & 0.74 & 0\\ 0 & 0.81 & 0.56 & 0.57 & 0.32 & 0 & 0.54 & 0 & 0.66 & 0.45\\ 0 & 0.75 & 0 & 0 & 0.35 & 0 & 0.46 & 0 & 0 & 0.4\\ 0 & 0.68 & 0.48 & 0.47 & 0.43 & 0 & 0.52 & 0 & 0.67 & 0.52\end{pmatrix}$$

По формуле (11) вычислим значения весовых коэффициентов абстрактного каталога K^{ab}, т.е. $\eta_{1j}^{ab}=\max\limits_{l=\overline{1,4}}\{\eta_{1j}^l\},\ j=\overline{1,10}$:

$$\{\eta_{1j}^{ab}\}=\{0\quad 0.81\quad 0.56\quad 0.57\quad 0.51\quad 0\quad 0.54\quad 0\quad 0.74\quad 0.52\}$$

Для всех тематических каталогов вычислим отклонения по формуле (12):

$$\lambda_{11}=\frac{5}{10}\cdot\sum_{j=1}^{10}\left(\eta_{1j}^{ab}-\eta_{1j}^1\right)^2=0.78,\qquad \lambda_{12}=\frac{6}{10}\cdot\sum_{j=1}^{10}\left(\eta_{1j}^{ab}-\eta_{1j}^2\right)^2=0.03,$$

$$\lambda_{13}=\frac{6}{10}\cdot\sum_{j=1}^{10}\left(\eta_{1j}^{ab}-\eta_{1j}^3\right)^2=0.74,\qquad \lambda_{14}=\frac{7}{10}\cdot\sum_{j=1}^{10}\left(\eta_{1j}^{ab}-\eta_{1j}^4\right)^2=0.04.$$

Наименьшее отклонение от K^{ab} имеет тематический каталог K_2, т.е. $\lambda_1^*=\min\limits_{l=\overline{1,4}}\{\lambda_{1l}\}=\min\{\lambda_{11},\lambda_{12},\lambda_{13},\lambda_{14}\}=\lambda_{12}=0.03$. Это означает, что тематический каталог K_2 является наиболее подходящим тематическим каталогом и информационный каталог d необходимо включить в него.

Если пороговое значение для отклонений тематик каталогов от абстрактного тематического каталога взять $\delta_1=\delta_2=\delta_3=\delta_4=0.1$, тогда тематические каталоги K_2 и K_4 являются наиболее релевантными, так как:

$\lambda_{12} < \lambda_{14} \le 0.1$. Соответственно, информационный ресурс d нужно включить как в тематический каталог K_2, так и в K_4.

4. Методы повышения точности выбора профиля и улучшения качества тематического каталога

В предыдущем разделе отметили, что для улучшения полноты и точности поиска можно использовать множество синонимов, ассоциирующих слов и словарей. Также было отмечено, что синонимы и ассоциирующие слова терминов объединяются в одном множестве $S=\{s_v\}_V=\{S_j\}_J$, где S_j - множество синонимов и ассоциирующих термина t_j. Для удобства в дальнейшем множество S назовем множеством синонимов. Необходимо отметить, что синонимы также являются терминами. Любое ключевое слово, которое является синонимом термина одного информационного ресурса, может оказаться важным (часто встречающимся) термином для другого. Это означает, что для любого термина $t_j \in T$ множество его синонимов S_j является подмножеством множества терминов T, т.е. $S_j \subset T$. Отсюда вытекает, что $S \subset T$. Для удобства в дальнейшем вместо множества S будем использовать множество T, таким образом, вместо отношения между t_j и s_v, рассмотрим отношение между t_j и t_v, где $j, v = \overline{1,J}$. Это отношение обозначим через r_{jv}^s, которое показывает степень близости t_v к t_j и представляется в виде нечеткой реляционной матрицы размерности JxJ:

$$r_{jv}^S = \left\{\nu_{jv} : T \times T \to [0,1]\right\}, j, v = \overline{1,J}, \quad R^S = \left\{r_{jv}^T\right\}_{J \times J} \qquad (15)$$

где ν_{jv} - функция принадлежности термина t_v в множество синонимов S_j, иначе говоря степень близости терминов t_j и t_v. Если для синонимов $t_j \in T$ и $t_v \in T$ не выполняется условие $t_v \in S_j$, тогда $\nu_{jv} = 0$.

Знание о синонимах и степени смысловой близости их к терминам дает возможность расширять знание о тематике информационного ресурса, что позволяет лучше определять тематический профиль и выбрать наиболее релевантный тематический каталог. Для улучшения (дополнения) знаний об информационном ресурсе нужно объединить знания о терминах и синонимах терминов. Если T_i^D - множество терминов информационного ресурса d_i и S_j - множество синонимов термина $t_j \in T_i^d$, $j = \overline{1,J}$, тогда

$$\tilde{T}_i^D = \bigcup_{t_j \in T_i^D} \left(S_j \cap T_i^D\right), \qquad (16)$$

является улучшенным множеством терминов данного информационного ресурса.

Значения новых отношений $\tilde{\mu}_{ij}$, соответствующие новому множеству $\tilde{T}_i^D$, можно найти следующим образом:

$$\tilde{\mu}_{ij} = \max_{v=\overline{1,J}} \left\{\nu_{jv} \cdot \mu_{iv}\right\}, j = \overline{1,J} \qquad (17)$$

Формула (17) дает нам новых (улучшенных) значений весовых коэффициентов всех терминов. Решение задачи (10)-(13) используя вместо μ_{ij}

новых весовых коэффициентов $\tilde{\mu}_{ij}$ позволяет достичь более эффективных результатов.

Пример 2. Возьмем те же множества T и K (также отношения R_l^K и R_l^D), которые определены в предыдущем примере. Пусть еще задана матрица отношений $R^S = \{r_{jv}^T\}_{J\times J}$:

$$R^S = \begin{pmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \\ S_7 \\ S_8 \\ S_9 \\ S_{10} \end{pmatrix} = \begin{pmatrix} 1 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.9 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.9 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.9 & 0 & 0 & 1 & 0.9 \\ 0 & 00 & 00 & 0 & 0 & 0 & 0 & 0 & 0.9 & 1 \end{pmatrix}$$

Тогда по формуле (17) получим значения :

$$R_1^D = (0.81 \quad 0.9 \quad 0.78 \quad 0.87 \quad 0.54 \quad 0.86 \quad 0.6 \quad 0 \quad 0.95 \quad 0.86)$$

Используя новые значения $\tilde{\mu}_{ij}$ вычислить по формуле (10)-(13) отклонения тематики каталогов, то получим:

$$\lambda_{11} = 0.94, \quad \lambda_{12} = 0.64, \quad \lambda_{13} = 1.1, \quad \lambda_{14} = 0.56.$$

Отсюда видно, что в отличии от предыдущего примера тематический каталог K_4 является наиболее релевантным данному документу, т.е. $\lambda_{14} = \min\limits_{l=\overline{1,4}} \lambda_{1l} = 0.56$. Следует отметить, что K_2 также является релевантным каталогом, коэффициент отклонения которого немного превышает λ_{14}.

5. Поиск наиболее релевантных информационных ресурсов по запросу пользователя

В информационно-поисковых системах Интернет поиск производится по запросу пользователя, который, как отметили выше, составляется из поисковых признаков - ключевых слов или терминов $Q = \{t_p\}_P$. В начале поиска в поисковой системе запросы преобразуются в соответствующий формат подобно формату представления информационных ресурсов. Для каждого признака определяется отношение его к данному запросу. Другими словами, указывается степень важности признаков для данного запроса, который представляется в виде нечеткого отношения:

$$r_p^Q = \{\alpha_p\} \to [0,1], \; p = \overline{1,P}, \quad R^Q = \{r_p^Q\}_P, \qquad (18)$$

где α_p - важность признака t_p для запроса Q.

Сначала необходимо нормировать множество Q, т.е. множество Q дополняется элементами множества T, которые отсутствуют в запросе, с

нулевыми коэффициентами важности. Тогда отношение (18) получит следующий вид:

$$r_p^{Q^T} = \left\{\alpha_p^T : T \times Q \to [0,1]\right\},\ p = \overline{1,J},\quad R^{Q^T} = \left\{r_p^{Q^T}\right\}_J, \tag{19}$$

Здесь если $t_p \in Q$, то $\alpha_p^T = \alpha_p$, иначе $\alpha_p^T = 0$.

Исходя соображений, из указанных в предыдущем разделе, для улучшения полноты и точности поиска будем использовать множество синонимов:

$$\widetilde{Q} = \bigcup_{t_j \in Q}\left(S_j \cap Q^T\right),\ j = \overline{1,J} \tag{20}$$

и, соответственно,

$$\widetilde{\alpha}_p = \max_{j=\overline{1,J}}\left\{v_{jp} \cdot \alpha_p^{Q^T}\right\},\ p = \overline{1,J}. \tag{21}$$

Теперь рассмотрим задачу поиска информации на основе запроса Q. Результатом поиска является список информационных ресурсов, релевантных к запросу пользователя, т.е. $IR = \left\{d_g\right\}_G$. Тогда

$$IR = Q \cap D = Q \cap \left(\bigcup_{i=\overline{1,I}} d_i\right) = \bigcup_{i=\overline{1,I}}\left(Q \cap d_i\right). \tag{22}$$

Задача (22) определения наиболее релевантных информационных ресурсов к запросу Q аналогична задаче (7) и для ее решения можно также использовать метод Беллмана-Заде..

Пересечение множеств отношений R_i^D и R^Q дает нам степени релевантности d_i к запросу Q, т.е. $R_i^{DQ} = R_i^D \cap R^Q$. Если $R_i^{DQ} = \left\{\gamma_{ij}\right\},\ i = \overline{1,I}\ \ j = \overline{1,J}$, тогда

$$\gamma_{ij} = \mu_{ij} \cdot \alpha_j,\ i = \overline{1,I},\ j = \overline{1,J}. \tag{23}$$

где γ_{ij} - степень релевантности документа d_i к запросу Q по отношению термина t_j.

Пусть d_i^{ab} - абстрактный информационный ресурс, который является идеально релевантным к запросу Q, тогда для степени релевантности d_i^{ab} к запросу Q по отношению термина t_j напишем:

$$\gamma_j^{ab} = \max_{i=\overline{1,I}}\left\{\gamma_{ij}\right\},\ j = \overline{1,J}, \tag{24}$$

Теперь определим суммарное среднеквадратическое отклонение $\gamma_{ij}, i = \overline{1,I}, j = \overline{1,J}$ от коэффициентов $\gamma_j^{ab}, j = \overline{1,J}$:

$$\lambda_i = \frac{\overline{J}}{J}\sum_{j=1}^{J}\left(\gamma_j^{ab} - \gamma_{ij}\right)^2, i = \overline{1,I}. \tag{25}$$

Информационный ресурс с минимальным отклонением от "идеального информационного ресурса" является наиболее релевантным:

$$\lambda^* = \min_{i=\overline{1,I}}\left\{\lambda_i\right\}. \tag{26}$$

Решение (23)-(26) позволяет определять один наиболее релевантный информационный ресурс d^*. Однако на практике требуется найти не одного источника, а всех наиболее релевантных источников информации.

Для решения данной проблемы введем параметр ε_i - пороговое значение релевантности. Тогда все информационные ресурсы, удовлетворяющие условия

$$\lambda_i \le \varepsilon_i,\ i=\overline{1,I} \qquad (27)$$

будем считать релевантными к запросу.

Пример 3. Пусть $Q=\{t_3/0.95, t_5/0.8, t_9/0.8\}$ - пользовательский запрос, а $D=\{d_1,d_2,d_3,d_4,d_5\}$ - документы информационного пространства, которые индексировались поисковой системой. В результате индексирования выявлены отношения терминов множества $T=\{t_1,t_2,t_3,t_4,t_5,t_6,t_7,t_8,t_9,t_{10}\}$ к документам d_j.

$$d_1(t_j/\mu_{1j})=\{t_1/0.7; t_2/0; t_3/0.9; t_4/0.8; t_5/0; t_6/0.7; t_7/0; t_8/0.8; t_9/0; t_{10}/0.9\},$$

$$d_2(t_j/\mu_{2j})=\{t_1/0.6; t_2/0; t_3/0.8; t_4/0; t_5/0.9; t_6/0.9; t_7/0; t_8/0; t_9/0.8; t_{10}/0\},$$

$$d_3(t_j/\mu_{3j})=\{t_1/0; t_2/0.6; t_3/0.9; t_4/0; t_5/0.9; t_6/0; t_7/0; t_8/0; t_9/0.7; t_{10}/0\},$$

$$d_4(t_j/\mu_{4j})=\{t_1/0; t_2/0; t_3/0; t_4/0; t_5/0.9; t_6/0.7; t_7/0.8; t_8/0; t_9/0.6; t_{10}/0\},$$

$$d_5(t_j/\mu_{5j})=\{t_1/0.6; t_2/0.7; t_3/0.9; t_4/0; t_5/0; t_6/0; t_7/0.6; t_8/0; t_9/0; t_{10}/0.8\}.$$

Множество Q дополним элементами T:

$$Q=\{t_1/0, t_2/0, t_3/0.95, t_4/0, t_5/0.8, t_6/0, t_7/0, t_8/0, t_9/0.8, t_{10}/0\}.$$

По формуле (23) найдем γ_{ij}:

$$\{\gamma_{ij}\}=\begin{pmatrix} 0 & 0 & 0.86 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0.76 & 0 & 0.72 & 0 & 0 & 0 & 0.64 & 0\\ 0 & 0 & 0.86 & 0 & 0.72 & 0 & 0 & 0 & 0.56 & 0\\ 0 & 0 & 0 & 0 & 0.72 & 0 & 0 & 0 & 0.48 & 0\\ 0 & 0 & 0.86 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Тогда из (24) вычислим γ_j^{ab}, $j=\overline{1,J}$:

$$\{\gamma_j^{ab}\}=(0\quad 0\quad 0.86\quad 0\quad 0.72\quad 0\quad 0\quad 0\quad 0.64\quad 0).$$

Из (25) и (26) получим:

$$\{\lambda_i\}=\begin{bmatrix} 0.82\\ 0.05\\ 0.06\\ 0.41\\ 0.68 \end{bmatrix} \text{ и } \lambda^*=0.05,$$

т.е. d_2 является наиболее релевантным к данному запросу. Если пороговые значения $\varepsilon_i=0.1$ для всех $i=\overline{1,5}$, тогда d_2 и d_3 являются релевантными.

Литература

[1] Корнеев В.В., Гарев А.Ф., Васютин С.В., Райх В.В. Базы данных: интеллектуальная обработка информации. -М.: "Нолидж", 2000. -352 с.

[2] Храмцов П. Моделирование и анализ работы информационно-поисковых систем Internet. Открытые Системы. Москва. № 6, 1996.

[3] Касумов В.А. Организация системы поиска в Азербайджанской части Internet. Москва. Журнал "Открытые системы". № 3, 2000, стр.59-62.

[4] Солтон Дж. Динамические библиотечно-информационные системы. - М.: Мир. 1979, 558 с.

[5] Касумов В.А. Поисковые механизмы библиотечно-информационных систем Internet. VI Международная конференция "Крым-2000". Библиотеки и ассоциации в меняющем мире, новые технологии и новые формы сотрудничества". Судак. Автономная Республика Крым. Украина. 3-11 июня 2000. стр.240-244.

[6] Yanhong Li. Toward a qualitative search engine. IEEE Internet Computing. July-August. 1998. pp.24-29. Internet: Http://coputer.org/internet/.

[7] Gudivada V.N. Information search on *World Wide Web*. Compouter Weekly, Moscow. № 35, 1997, pp. 19-21, 26,27.

[8] Касумов В.А. Разработка полнотекстовой поисковой системы по информационным ресурсам Азербайджана. VI Международная научно-практическая конференция "Проблемы создания, интеграции и использования научно-технической информации на современном этапе". Киев. 16-17 декабря 1999. стр.15-16.

[9] Roberto Okada, Eun-Seok Lee, Tetsuo Kinoshita, Nurio Shiratori. A method for personalized web searching with hierarchical document clustering. Transaction of Information Processing Society of Japan. Vol.39, N:4, Apr. 1998, pp. 868-877.

[10] Орловский С.А. Проблемы принятия решений при нечеткой исходной информации. М.: Наука. 1981. 208 с.

THE MODELLING OF INFORMATION RETRIEVAL IN INTERNET ON BASE OF FUZZY KNOWLEDGE

V.A.GASIMOV

Information retrieval systems of the Internet created on the base of the thematic catalogue, give higher precise results. Despite of it, they have a few disadvantages, so for example, the creation and management of the thematic catalogue is executed by the person manually, and it requires huge intellectual activity and professionalism of staff, the updating of the database of such system is very laborious and long-continued process, with increase of quantity of subdirectories in the thematic catalogue complicates search process of the necessary information sources and obtaining of desired results. The retrieval systems based on automatic indexes, on the contrary, are more flexible and simple for implementation, cover more information sources, than thematic catalogues, i.e. the recall of automatic indexes much more exceeds recall of the thematic catalogues. However precision of search results of automatic index is lower than precision of the thematic catalogues.

The researches show, that the development of methods of creation of the automatic thematic catalogues without person's interference is actual and much promising. Here it is supposed, that the search robot of a system itself should determine subjects of information resource, the most eligible theme profiles from the thematic catalogues and to include resource into them. Thus the created thematic catalogues can be built according to present standards УДК, ББК etc. The created in this way automatic thematic catalogue can substitute the person (manager or operator of the retrieval system), will allow to execute activities on creation, service and updating of the thematic catalogue of the retrieval system.

In the present article the model of information search is researched. The methods of definition of information resources subjects, splitting of the information space on thematic directions and improvement of the thematic catalogue are considered. For the solution of formulated problems the theory of fuzzy sets and relations, and also approach Bellman-Zadeh are used.

Ключевые слова: информационный поиск, поисковая система, информационный ресурс, тематический каталог, автоматический индекс, автоматический тематический каталог, нечеткие множества, нечеткие отношения.

Keywords: information retrieval, retrieval system, information resource, thematic catalogue, automatic index, automatic thematic catalogue, fuzzy sets, fuzzy relations.