УДК 681.3.06

Э.М. Котов, А.Н. Целых

ИССЛЕДОВАНИЕ МОДЕЛЕЙ ИНФОРМАЦИОННОГО ПОИСКА

При рассмотрении моделей информационного поиска возможно говорить о нескольких моделях, начиная с классических, таких как пространственно-векторные, вероятностные и булевы модели. В подобных моделях, моделируются средства поиска, релевантные объекты информационного поиска (главным образом документы, запросы и термины) их связи друг с другом и как они организованы в структуры, используя векторы, вероятности или логические операторы. В статье рассматривают некоторые формальные методы, подчеркнуто их богатство при использовании в качестве модели для информационного поиска.

Информационный поиск; модели и методы информационного поиска.

E.M. Kotov, A.N. Tzelykh

RESEARCH OF MODELS FOR INFORMATION RETRIEVAL

By consideration of models of information retrieval probably to speak about several models, since classical, such as Vector Space models, Probabilistic models and Boolean models. In similar models, search means, relevant objects of information retrieval (documents, inquiries and terms) their communications with each other and as they are organised in structures, using vectors, probabilities or logic operators are modelled. In article some formal methods are considered, their riches are underlined at use in quality the general models for information retrieval. It has allowed us to identify components and possible relations.

The information retrieval; models and methods of information retrieval.

Формализация и модель − те термины, которые должны использоваться с осторожностью применительно к системам Информационного поиска (ИП). Возможно говорить о нескольких моделях, начиная с классических, таких как пространственно-векторные предложенные Джеральдом Салтоном [1], вероятностные и булевы модели. В подобных моделях, моделируются средства поиска, релевантные объекты ИП (главным образом документы, запросы и термины) их связи друг с другом и как они организованы в структуры, используя векторы, вероятности или логические операторы. Кроме того, если исследовать другие подходы для ИП как семантическая индексация [2], нейронные сети [3] или генетические алгоритмы [4], возможно видеть, что они так же часто упоминаются как «модели», хотя их целесообразней назвать поисковыми стратегиями [5]. Формальная модель состоит в представлении, используемом для детализации поисковой стратегии. Таким образом, можно сказать: Формальная модель для информационного поиска есть математическое представление, способное отобразить любой релевантный объект в информационно-поисковой системе, наряду с любыми отношениями (функциями, картами, бинарными отношениями и т.д.) для использования системой, с целью выполнения поисковой задачи.

Различия между существующими поисковыми стратегиями породили широкое разнообразие моделей. Модель состоит в представлении, используемом для детализации поисковой стратегии. Таким образом, можно сказать, что формальная модель для информационного поиска есть математическое представление, способное отобразить любой релевантный объект в информационно-поисковой системе, наряду с любыми отношениями (функциями, картами, бинарными отношениями и

т.д.) для использования системой, с целью выполнения поисковой задачи. Если модель будет достаточно общая, то это будет полезно только для очень поверхностной концептуализации задачи информационного поиска. С другой стороны, если модель определена достаточно глубоко, чтобы охватить все возможные аспекты системы, то возникает проблема в сложном описании, что создает трудности для дальнейшего расширения вместо того, чтобы быть практическим и дорабатываемым.

Можно классифицировать модели, доступные в литературе, в зависимости от выбранной математической основы на логические и алгебраические. В качестве примера, иллюстрирующего общую алгебраическую модель, выберем вариант, предложенный в [5]. Они определяют систему ИП как кортеж:

$$I = (D, Q, \delta),$$

где D – множество документов; Q – множество запросов; δ – поисковая функция.

$$\delta : Q \to 2^D, q \mapsto \delta(q) := \delta_i \in 2^D,$$

где 2^D – является множеством всех возможных подмножеств D.

Следовательно, поисковая функция δ производит подмножество документов δ_i как ответ на вопрос $q_i \in Q$. Эта модель может быть легко расширена, чтобы включить в неё тезаурус или описать распределенный ИП. С тезаурусом мы имеем:

$$I = (T, D, Q, \delta),$$

где T – множество различных терминов (управляемый словарь) с отношениями: $p \subset T \times T$ таким, что $p(t_1, t_2)$ подразумевает, что термины t_1 и t_2 являются синонимами.

Эти отношения дают разделение множества T на подмножества синонимов, то есть все термины в подмножестве – синонимы.

Другой пример, немного более детальный, это модель, предложенная Шериданом. Используется кортеж:

$$\langle T, \Phi, D; ff, df \rangle,$$

где T – множество возможных терминов в документе; Ф – является множеством особенностей индексации (лемматизации, выбора стоп-листов):

$$\phi : T \to \Phi, \tau \mapsto \phi(\tau) := \phi_i;$$

где D – множество документов: $\phi : T \to D, \tau \mapsto d(\tau) := d_j, \tau \in T, \phi_i \in \Phi$,

ϕ_i – лемматизированная версия термина τ;

$$ff(\phi_i, d_j) = |\{\tau \in T \mid \phi(\tau) = \phi_i \wedge d(\tau) = d_j\}|;$$

$$df(\phi_i) = |\{d_j \in D \mid \exists \tau \in T : \phi(\tau) = phi_i \wedge d(\tau) = d_j\}|.$$

Поисковые стратегии, примененные в данной структуре, делают вполне законченной модель ИП. Эта структура весьма интересна, но отсутствует компонент, который фактически соотносит вопросы с документами, и таким образом нельзя полностью квалифицировать эту структуру как модель для системы ИП. Рассмотрим модель, предложенную Бэйза-Ятес [6]. Эта модель более интересна, чем пред-

ставленные выше, в связи с тем, что использует функцию ранжирования, и, таким образом, модель практически ближе к текущим поисковым стратегиям:

Информационно-поисковая модель представлена как четверка:

$$\langle D, Q, F, R \rangle,$$

где D – является множеством представлений документа; Q – множество запросов; F – структура моделирующая документы, запросы и их отношения; R – функция ранжирования:

$$R: Q \times D \to \Re, \langle q_i, d_j \rangle \mapsto R(q_i, d_j) := r_{ij} \in \Re.$$

Гибкость модели заключается в структуре компонентов. Это может быть векторное пространство с его операторами, алгебра для булевой модели, или любая другая структура способная моделировать поисковую стратегию. Данная модель является полной, но слишком общей с практической точки зрения.

Обширная работа над формализацией моделей ИП проведена Сандором Доминичем [7,8]. Он предлагает приемлемую структуру для любой классической модели информационного поиска (векторной, вероятностной и булевой моделей). Рассмотрим описание этой модели.

Прежде всего, чтобы разъяснять формализм, используемый позже, определим следующие понятия.

Идентификаторы – любая часть информации, используемой для описания документа (термины, индексы, ключевые слова, описатели и т.д.).

Объекты – любая информация, подходящая, для составления документа (текст, изображения, звуковые фрагменты и т.д.). Может быть, конечно, непосредственно документ.

Документ – группа объектов. Во многих случаях, когда коллекция состоит только из полнотекстовых документов, документ содержит только один объект: его текст. По этой причине ряд моделей могут объединять эти два элемента в один.

Критерии – отражают взвешенные отношения между двумя документами (например: подобие, релевантность, расстояние и т.д.).

Порог – этот компонент используется для отбора множества документов, удовлетворяющих критерию выше порогового значения.

Поиск – является отображением документа к множеству документов.

Дадим формальное описание:

1. $T = \{t_1, t_2, ..., t_k, ..., t_N\}$ – множество идентификаторов $N \geq 1$.
2. $O = \{o_1, o_2, ..., o_u, ..., o_U\}$ – множество объектов $U \geq 1$.
3. $(D_j)_{j \in J = \{1,2,...,M\}}$ – множество групп объектов $D_j \in 2^O M \geq 1$.
4. $D = \{\tilde{o}_j \mid j \in J\}$ – ряд документов, где нечеткое множество $\tilde{o}_j = \{(t_k, \mu_{\tilde{o}_j}(t_k)) \mid t_k \in T, k = 1,2,...,N\}, j = 1,2,...,M$, $\mu_{\tilde{o}_j} : T \to S \subseteq [0,1] \subset R$ является представителем группы объектов D_j.
5. $A = \{\tilde{a}_1, \tilde{a}_2, ..., \tilde{a}_i, ..., \tilde{a}_C\}$ – множество критериев $C \geq 1$, где $\tilde{a}_i = \{((q, \tilde{o}_k), \mu_{\tilde{a}_i}(q, \tilde{o}_k)) \mid \tilde{o}_j \in D, j = 1,2,...,M\}, i = 1,2,...,C$, является нормализованным нечетким отношением и $\mu_{\tilde{a}_i} : D \times D \to [0,1] \subset R$.

6. $a_{\alpha_i} = \{\tilde{o} \in D \mid \mu_{\tilde{a}}(q,\tilde{o}) > \alpha_i\}, i = 1,2,...,C$, – α_i – порог критерия $\tilde{a}_i$, $0 \le \tilde{\alpha}_i < \infty$

7. $\Re : D \to 2^D$ – является отображением называемым поиском.

Классический Информационный поиск (КИП) определяется как система, сформированная множеством документов и функцией поиска, отображающаяся в 2 кортежах:

$$\langle D, \Re \rangle .$$

со следующими свойствами:

Свойство 1. $q = \tilde{o} \Rightarrow \mu_{\tilde{a}_i}(q,\tilde{o}) = 1, \forall q,\tilde{o} \in D, i = 1,2,...,C$ – свойство рефлективности.

Свойство 2. $\Re(q) = \{\tilde{o} \mid \mu_{\tilde{a}_i}(q,\tilde{o}) = \max_{k=1,...,C} \mu_{\tilde{a}_i}(q,\tilde{o})\} \bigcap a_{\alpha_i}$,

i – выбирается произвольно.

Первая свойство характеризует то, что в случае соответствия документа запросу, любой критерий должен возвратиться со значением 1.

Второе свойство устанавливает один произвольный критерий и поиск будет являться пересечением между двумя множествами: множество документов с весом, установленным в соответствии с критерием по выбранному порогу (α_i) и множеством документов, которые имеют вес с данным критерием всегда выше чем вес, возвращенный в соответствии с любыми другими критериями.

Графическое представление второго свойства может быть продемонстрировано на рис 1.

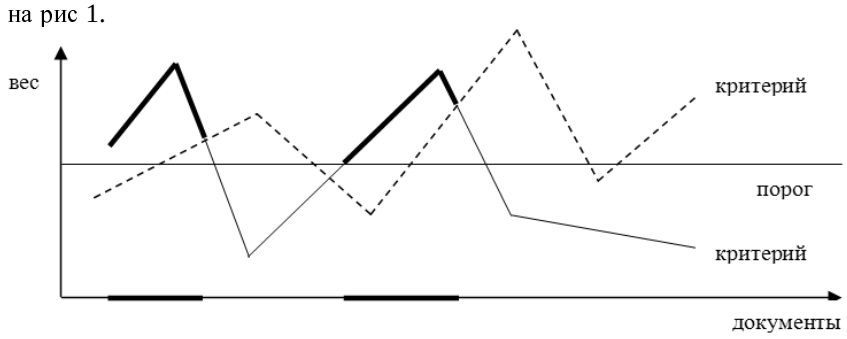

Рис. 1. Релевантность в классической ИПС

Здесь, для некоторого запроса, используя критерий 2, были бы отобраны документы, выделенные жирной линией показанной на рисунке (см. рис 1). Используя этот подход, легко определить пространственно-векторные и вероятностные модели.

$$IR = m[\Re(O,(Q,\langle I,\delta \rangle))],$$

где O – множество объектов (документы); Q – множество запросов; I – пользовательская информация; δ – является информацией, получаемой из пользовательской информации I, получаемая по определенным правилам; $\Re$ – является отношением между объектами и информационной потребностью.

Информационная потребность, выражается:

$$IN = (Q,\langle I,\delta \rangle) .$$

Данная модель формализует, так называемую, пользовательскую информацию, так как здесь учитывается персонифицированная информация о пользователе,

чтобы ввести дополнительную информацию при удовлетворении его информационная потребность. Языковая грамматика используется как средство представления документов и запросов к нормальной форме. Следовательно, и те, и другие могут быть представлены как булево выражение, составленное в соответствии с условиями и логическими операторами $(\wedge, \vee, \neg)$.

В итоге отметим, что были рассмотрены некоторые формальные методы, подчеркнуто их богатство при использовании в качестве модели для ИП. Это позволило нам идентифицировать общие компоненты и возможные отношения.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Salton G.* A Theory of Indexing, Technical report No. TR74-203, Department of Computer Science, Cornell University, Ithaca, New York, 1974.
2. *Bartell Brian T., Cottrell Garrison W., Belew Richard K.* Latent Semantic Indexing is an Optimal Special Case of Multidimensional Scaling. In: Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. P. 161-167, 1992.
3. *Belew Richard K.* Adaptive Information Retrieval: Using a Connectionist Representation to Retrieve and Learn about Documents. In: Proceedings of the Twelfth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. P. 11-20, 1989.
4. *Chen X.* Cold Accretion Disks with Coronae and Advection. Astrophysical Journal v. 448, P. 803, 1995.
5. *David A. Grossman, Ophir Frieder, M. Catherine McCabe, Abdur Chowdhury.* A unified environment for fusion of information retrieval approaches. Conference on Information and Knowledge Management. Kansas City, Missouri, United States. Pages: 330 – 334, 1999.
6. *Baeza-Yates R.* Modern Information Retrieval / R. Baeza-Yates, B. Ribeiro-Neto. – New York: ACM Press Series/Addison Wesley, 1999.
7. *Sandor Dominich*. A unified mathematical definition of classical information retrieval. JASIS 51(7) Pages: 614-624, 2000.
8. *Sandor Dominich, Mounia Lalmas, C. J. van Rijsbergen.* ACM SIGIR 2000 Workshop on Mathematical/Formal Methods in Information Retrieval. SIGIR Forum 34(1) Pages: 18-23, 2000.

Целых Александр Николаевич
Технологический институт федерального государственного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге.
E-mail: info@tti.sfedu.ru.
347928, г. Таганрог, пер. Некрасовский, 44.
Тел.: 8(8634)371-160.
Заместитель руководителя по информатике.

Котов Эдуард Михайлович
Технологический институт федерального государственного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге.
E-mail: kotov@tti.sfedu.ru.
347928, г. Таганрог, пер. Некрасовский, 44.
Тел.: 8(8634)371-743.
Кафедра прикладной информатики; старший преподаватель.

Tselykh Aleksandr Nikolaevich
Taganrog Institute of Technology – Federal State-Owned Educational Establishment of Higher Vocational Education “Southern Federal University”.
E-mail: info@tti.sfedu.ru.
44, Nekrasovskiy, Taganrog, 347928, Russia.

Phone: 8(8634)371-160.
Vice Rector for Informatics.

Kotov Eduard Mihailovich
Taganrog Institute of Technology – Federal State-Owned Educational Establishment of Higher Vocational Education “Southern Federal University”.
E-mail: kotov@tti.sfedu.ru.
44, Nekrasovskiy, Taganrog, 347928, Russia.
Phone: 8(8634)371-743.
Department of Applied Information Science; senior instructor.

УДК 621.396

Ю.М. Вишняков, С.Ю. Новиков

О ПОДХОДЕ К УПРАВЛЕНИЮ УРОВНЕМ СЕРВИСОВ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

В работе рассматривается задача управления уровнем услуг информационных систем предприятий. Представлен тривиальный алгоритм управления, обладающий гарантированной сходимостью. На его основе разработаны алгоритмы управления по отклонению, возмущению, а также комбинированный алгоритм. Для сокращения погрешности, неизбежно возникающей при управлении реальными объектами, был предложен метод последовательного сокращения промежутка исследуемых данных.

На основе предложенных методов может быть построена модель сервисно-ориентированной информационной системы и автоматизирован процесс управления информационными системами предприятий.

Уровень сервиса; SLA; алгоритмы управления уровнем сервиса.

Y.M. Vishnyakov, S.Y. Novikov

THE APPROACH TO THE MANAGEMENT LEVEL IN INFORMATION SYSTEMS

In this paper we consider the problem of management of information systems services to businesses. A trivial algorithm, which has guaranteed convergence, was presented. On the basis of trivial algorithm developed algorithms for the rejection, indignation, and the combined algorithm. To reduce errors, has been proposed a method of reducing the gap investigated serial data.

On the basis of the proposed methods can be created a model of service-oriented information system and built automated management of information systems.

Level of service; service level agreement; algorithms for managing service level.

Перед современным промышленным предприятием сегодня стоит задача обеспечения уровня предоставляемых услуг в области информационных технологий (ИТ-сервисов) на требуемом для бизнеса уровне. Требования бизнеса к уровню услуг, предоставляемых подразделением, ответственным за развитие предприятия в области информационных технологий (ИТ-подразделением) изложены в формализованном виде в специальном документе, называемом соглашением об уровне сервиса (Service Level Agreement – SLA) [1]. Поскольку в настоящее время пользователям предоставляется большое число различных сервисов, то задача обеспечения уровня указанного в SLA представляет собой довольно сложную задачу, не всегда имеющее однозначное решение. Множество контролируемых параметров,