УДК 025.4:681.3

ПОИСК И РАНЖИРОВАНИЕ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ МУЛЬТИАГЕНТНОЙ СИСТЕМЫ

Иванова Г.С., Андреев А.М., Шоуман М.А.

*Московский государственный технический университет
им. Н.Э. Баумана, Москва, e-mail: marwashouman834@yahoo.com*

Информационный поиск – деятельность, направленная на получение информационных ресурсов, релевантных информационным потребностям пользователя, из имеющейся в Интернете коллекции. Количество документов, содержащих требуемую информацию, в том числе в неструктурном виде, непрерывно растет. Соответственно растет и количество документов, находимых по поисковым запросам, что, в свою очередь, приводит к тому, что задача оценки релевантности документов запросу все еще является актуальной. В статье рассматривается модель структуры мультиагентной поисковой системы в виде нечеткого метаграфа. На основе этой модели получена система нечеткого логического вывода, позволяющая ранжировать найденные системой документы по степени их релевантности. Приведены результаты двух экспериментов по оценке релевантности результатов поиска с применением систем нечеткого вывода Мамдани и Сугено.

Ключевые слова: информационный поиск, мультиагентные системы, модель векторного пространства, нечеткий метаграф, системы нечеткого вывода

RETRIEVAL AND RANKING OF DOCUMENTS USING MULTI-AGENT SYSTEM

Ivanova G.S., Andreev A.M., Shouman M.A.

Bauman Moscow State Technical University, Moscow, e-mail: marwashouman834@yahoo.com

Information retrieval is the activity of obtaining information resources relevant to an information need of the user from a collection of information resources available on the Internet. The amounts of the documents contain required information, including the unstructured form continuously growing. Consequently it is increasing the number of documents being retrieved on search queries, which leads to the fact that the task of evaluating the relevance of the documents to the query is still relevant. In this article the Fuzzy Metagraph structure model of the Multi-agent information retrieval system is proposed. Based on the proposed model a fuzzy logic inference system is obtained providing ranking of the retrieved results according to their relevance. The evaluations of the Multi-agent System results depend on the comparison of fuzzy inference types Mamdani and Sugeno.

Keywords: information retrieval, multi-agent system, vector space model, fuzzy metagraph, fuzzy inference system

Основная задача информационного поиска – нахождение максимального количества документов, релевантных информационным потребностям пользователя. Однако количество документов, содержащих ключевые слова, может быть велико и будет увеличиваться с увеличением общего количества документов в сети Интернет. Следовательно, ранжирование документов по степени их релевантности является актуальным сейчас и сохранит свою актуальность в будущем.

Подавляющее множество документов в сети содержит неструктурированную информацию, а потому выявление степени соответствия документа запросу – задача нетривиальная и предполагает использование интеллектуальных методов. В настоящее время разработано большое количество методов, в том числе интеллектуальных, оценки релевантности результатов информационного поиска [4, 5]. Однако эти методы имеют большую вычислительную сложность, а потому малоприменимы при большом количестве полученных результатов.

В настоящей работе для реализации системы используется мультиагентный подход, при котором запрашиваемые действия выполняются некоторым количеством отдельно функционирующих агентов. Подход обеспечивает не только снижение сложности системы, но и возможность распараллеливания процесса поиска для сокращения времени обработки запроса.

Целью настоящего исследования является ранжирование извлеченных неструктурных документов по увеличению их релевантности исходному поисковому запросу в рамках мультиагентной интеллектуальной системы информационного поиска.

Построение модели мультиагентной интеллектуальной системы информационного поиска

Мультиагентный подход эффективен в том случае, если структура системы и взаимодействие между объектами хорошо проработаны. Рассмотрим функции, которые должна выполнять интеллектуальная система информационного поиска.

Система, предназначенная для осуществления автоматического информационного поиска, должна реализовать пять основных этапов обработки:

- ввод ключевых слов;
- поиск в Интернете по ключевому слову;
- извлечение требуемой информации из Веб-источников;
- интеллектуальный анализ добытых текстов;
- сохранение выходных данных в базе данных [1].

Эти действия могут осуществляться пятью агентами: интерфейсным, поисковым, агентом извлечения информации, агентом интеллектуальной обработки текстов и агентом ранжирования документов. В соответствии с мультиагентной технологией все агенты общаются друг с другом с помощью сообщений. Использование ограниченного количества агентов, выполняющих не сильно связанные операции, позволяет сократить количество передаваемой между ними информации и, следовательно, уменьшить суммарное время поиска релевантных запросу документов.

Первый агент – интерфейсный – обеспечивает интерфейс пользователя, позволяющий взаимодействовать с системой через графические тексто-ориентированные интерфейсы путем ввода ключевого слова.

Второй агент – поисковый – отправляет ключевое слово в поисковую машину Google, которая возвращает ссылки, собирая URL доступных веб-сайтов в Интернете, и передает их третьему агенту.

Третий агент – извлечения информации – автоматически извлекает тексты по URL-ссылкам, в том числе большое количество различных неструктурированных текстовых ресурсов, и передает их для дальнейшей обработки.

Четвертый агент – интеллектуальной обработки текстов – извлекает полезную информацию из текста, используя лексемизацию (удаление знаков препинания, специальных символов и замену отступов и других нетекстовых символов одним пробелом) и фильтрацию стоповых слов (удаление слов, которые не относятся к документам). А также вычисляет веса терминов в документах (TF-IDF) [1] для дальнейшей оценки степени релевантности документа.

И, наконец, пятый агент выполняет операцию ранжирования документов, используя для оценки соответствия документа запросу системы нечеткого вывода Мамдани и Сугено.

Поскольку агент может выполнять более чем одну операцию, модель, представляющая процесс поиска, должна быть ие-

рархической (двухуровневой). В качестве такой модели будем использовать нечеткий метаграф, который также позволит отобразить нечеткую характеристику соответствия документа запросу.

Метаграф представляет собой иерархическую структуру, основанную на графе, в которой каждый узел является множеством, имеющим один или более элементов. Эта модель сохраняет все свойства графов [2–3, 6–7].

Структура нечеткого метаграфа $\tilde{S}$, представляющего собой модель процесса поиска, описывается следующим образом:

$$\tilde{S} = \{X, \tilde{X}, \tilde{E}\},$$

где $X = \{x_i, i = \overline{1, 10}\}$ – множество операций, осуществляемых в процессе поиска и интеллектуальной обработки документов; $\tilde{X}$ – нечеткое множество на X – множество операций, осуществляемых в процессе поиска и интеллектуальной обработки документов с учетом неопределенности $\tilde{x}_i = \{x_i, \mu(x_i)\}$, где μ – функция принадлежности, $(\forall x_i \in X) \mu: x_i \rightarrow [0, 1]$ (таблица); $\tilde{E} = \{\tilde{e}_1, \tilde{e}_2, \tilde{e}_3, \tilde{e}_4\}$ – множество сообщений, передаваемых между агентами:

$$\tilde{e}_1 = \left\{ \left\{ \tilde{x}_1 \right\}, \left\{ \tilde{x}_2, \tilde{x}_3, \tilde{x}_4 \right\} \right\};$$

$$\tilde{e}_2 = \left\{ \left\{ \tilde{x}_2, \tilde{x}_3, \tilde{x}_4 \right\}, \left\{ \tilde{x}_5 \right\} \right\};$$

$$\tilde{e}_3 = \left\{ \left\{ \tilde{x}_5 \right\}, \left\{ \tilde{x}_6, \tilde{x}_7, \tilde{x}_8, \tilde{x}_9 \right\} \right\};$$

$$\tilde{e}_4 = \left\{ \left\{ \tilde{x}_6, \tilde{x}_7, \tilde{x}_8, \tilde{x}_9 \right\}, \left\{ \tilde{x}_{10} \right\} \right\}.$$

Каждому агенту в модели соответствует подмножество операций, которые он выполняет:

- интерфейсному агенту – $\tilde{X}_1 = \{\tilde{x}_1\}$;
- поисковому агенту – $\tilde{X}_2 = \{\tilde{x}_2, \tilde{x}_3, \tilde{x}_4\}$;
- агенту извлечения документов – $\tilde{X}_3 = \{\tilde{x}_5\}$;
- агенту интеллектуальной обработки текстов – $\tilde{X}_4 = \{\tilde{x}_6, \tilde{x}_7, \tilde{x}_8, \tilde{x}_9\}$;
- агенту ранжирования результатов поиска – $\tilde{X}_5 = \{\tilde{x}_{10}\}$.

Полученный нечеткий метаграф мультиагентной системы информационного поиска показан на рис. 1.

Условные обозначения элементарных операций поиска

Элемент множества	Обозначение	Моделируемая операция
$\tilde{x}_1$	UI	Ввод ключевого словосочетания через интерфейс пользователя
$\tilde{x}_2$	GIS	Передача ключевого сочетания <i>Google</i>
$\tilde{x}_3$	CURL	Получение ссылок на документы
$\tilde{x}_4$	SURL	Сохранение ссылок
$\tilde{x}_5$	RD	Извлечение документов по ссылке
$\tilde{x}_6$	DT	Лексемизация текста документа
$\tilde{x}_7$	DF	Фильтрация текста документа
$\tilde{x}_8$	DS	Выделение частей речи в тексте документа
$\tilde{x}_9$	TW	Вычисление веса каждого термина в документе
$\tilde{x}_{10}$	OE	Оценка соответствия документа ключевым словам

Мультиагентная система, построенная по указанной модели, включает систему принятия решения о степени соответствия документа запросу.

Система принятия решений о степени соответствия документа запросу

В основу системы принятия решения о степени соответствия документа запросу, включенной в агент 5, положены системы нечеткого вывода Мамдани и Сугено. Исходными данными для систем нечеткого вывода являются оценки весов соответствия для каждого термина запроса и для каждого документа из найденных [1], рас-

считанные агентом 4. Применение при этом интеллектуальной обработки текстов позволяет получить более точные оценки весов терминов.

Правила для систем нечеткого вывода автоматически формируются в процессе поиска с использованием модели процесса интеллектуального поиска – нечеткого метаграфа. Количество правил зависит от количества терминов в запросе.

Для получения интегрированной оценки степени релевантности документа d_i запросу q в эксперименте используем косинусную меру оценки сходства, описанную в [5]:

$$\cos(\vec{q}, \vec{d}_i) = \frac{\vec{q} \cdot \vec{d}_i}{|\vec{q}| \cdot |\vec{d}_i|} = \frac{\sum_{j=1}^m w(t_j, d_i) \cdot w(t_j, q)}{\sqrt{\sum_{j=1}^m w^2(t_j, d_i)} \sqrt{\sum_{j=1}^m w^2(t_j, q)}},$$

где $w(t_j, d_i)$ – вес j -го термина в i -м документе d_i ; $w(t_j, q)$ – вес j -го термина в запросе q .

Рис. 1. Структура процесса интеллектуального информационного поиска в виде нечеткого метаграфа

Выбор этой меры обусловлен ее сравнительной простотой, что позволяет применять ее совместно с системами нечеткого вывода.

Экспериментальные исследования применения систем нечеткого вывода Мамдани и Сугено для получения оценок степени релевантности найденных документов запросу

Вычислительный эксперимент был проведен на наборе документов, извлеченных мультиагентной системой по ключевому словосочетанию «Компьютерные науки».

Таким образом, вектор запроса $q = (t_j)$, $j = 1, m$ включал два термина ($m = 2$):

$q = (t_1, t_2) = (\text{"Компьютерные"}, \text{"науки"}).$

С помощью мультиагентной системы было получено множество D документов (примерно 1,5 млн), содержащих заданные термины. Для каждого документа $d_i \in D$ в системе были определены веса терминов t_1 и $t_2 - w(t_1, d_i)$ и $w(t_2, d_i)$. Веса тех же терминов в запросе были приняты одинаковыми: $w(t_1, q) = 0,7$, $w(t_2, q) = 0,7$.

В системах нечеткого вывода Мамдани и Сугено были использованы правила:

1. Если $(w(t_1, q) - \text{высокое})$ и $(w(t_1, d_i) - \text{высокое})$, то $(\cos(q, d_i) - \text{высокая оценка})$.
2. Если $(w(t_1, q) - \text{низкое})$ и $(w(t_1, d_i) - \text{низкое})$, то $(\cos(q, d_i) - \text{низкая оценка})$.
3. Если $(w(t_2, q) - \text{высокое})$ и $(w(t_2, d_i) - \text{высокое})$, то $(\cos(q, d_i) - \text{высокая оценка})$.
4. Если $(w(t_2, q) - \text{низкое})$ и $(w(t_2, d_i) - \text{низкое})$, то $(\cos(q, d_i) - \text{низкая оценка})$.

Рис. 2. Функция принадлежности входа $w(t_2, q)$

Рис. 3. Результаты оценки соответствия документа запросу с использованием системы нечеткого вывода Мамдани:

а – оценка для наименее соответствующего документа; б – оценка для наиболее соответствующего документа; в–г – распределение оценок для 35 документов по терминам t_1 и t_2

В процессе нечеткого вывода типа Мамдани для входных лингвистических переменных $w(t_1, q)$, $w(t_2, q)$, $w(t_1, d)$ и $w(t_2, d)$ использовались треугольные функции принадлежности, как показано в примере на рис. 2 (форма функции показана в нотации пакета MATLAB).

В эксперименте с применением системы нечеткого вывода типа Мамдани веса терминов, полученные после интеллектуальной обработки текстов документов, пересчитывались с использованием треугольной функции принадлежности и приведенных выше правил. Далее были рассмотрены первые 35 документов. Расчет балла соответствия для документа, наименее соответствующего запросу из первых 35, показан на рис. 3, а (значение центроида 0,573).

Расчет балла соответствия для документа, наиболее соответствующего запросу

косинусного сходства для 35 документов по двум терминам «Компьютерная», «наука», показаны на рис. 3, в–г.

Во втором эксперименте применялась система нечеткого вывода Сугено. Также были рассмотрены первые 35 документов. Расчет балла соответствия для документа, наименее соответствующего запросу, показан на рис. 4, а (значение центроида 0,714). Расчет балла соответствия для документа, наиболее соответствующего запросу, показан на рис. 4, б (значение центроида 0,999). Оценки балла соответствия для остальных 33-х документов распределены в диапазоне значений равно 0,285, что примерно в 3 раза больше, чем в предыдущем эксперименте. Графически распределение результатов вычисления косинусного сходства для 35 документов по двум терминам «Компьютерная», «наука» показаны на рис. 4, в–г.

Рис. 4. Результаты оценки соответствия документа запросу с использованием системы нечеткого вывода Сугено:

а – оценка для наименее соответствующего документа; б – оценка для наиболее соответствующего документа; в–г – распределение оценок для 35 документов по терминам t_1 и t_2

су, показан на рис. 3, б (значение центроида 0,666). Оценки балла соответствия для остальных 33-х документов распределены в диапазоне значений равно 0,093. Графически распределение результатов вычисле-

Выполненные эксперименты позволяют сделать вывод, что для оценки балла ранжирования документа с использованием метода сходства (косинус) можно использовать оба типа систем нечеткого вывода. Однако

система нечеткого вывода Сугено обеспечивает лучшие результаты, чем система Мамдани, поскольку обеспечивает оценку, близкую к оценкам, получаемым аналитическими способами с существенно большей вычислительной сложностью [4, 5].

Выводы

Выполнена декомпозиция интеллектуальной поисковой системы для мультиагентной реализации по принципу выделения обобщенных операций, что позволяет сократить объем информации, передаваемой агентами в процессе работы. Для описания структуры поиска построен нечеткий метаграф, который учитывает неопределенность оценки соответствия результатов поиска введенным ключевым словам.

Для получения оценки соответствия документа запросу в условиях неопределенности исходных оценок соответствия предложено использовать системы нечеткого логического вывода Мамдани и Сугено и косинусную меру оценки сходства. Выполненные эксперименты позволяют сделать вывод, что для ранжирования найденных документов целесообразно использовать систему нечеткого вывода Сугено, которая обеспечивает оценку, близкую к оценкам, получаемым аналитическими способами, с существенно меньшей вычислительной сложностью.

Список литературы

1. Иванова Г.С. Автоматический поиск информации с использованием мульти-агентной системы / Г.С. Иванова, А.М. Андреев, В. И. Неведов, М.А. Шоуман, Е.В. Егорова // Электромагнитные волны и электронные системы. – 2015. – Т. 20 – № 2. – С. 33–38.
2. Иванова Г.С. Модели объектов задач структурного синтеза // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. – 2006. – № 12. – Режим доступа: <http://technomag.edu.ru/doc/62361.html> (дата обращения 15.09.2015).
3. Dashore P., Jain S. Fuzzy Metagraph and Hierarchical modeling // International Journal on Computer Science and Engineering. – 2011. – Vol. 3. – № 1. – P.435–449.
4. Raghavan P., Schütze H., Manning D. Introduction to Information Retrieval // Cambridge University Press. – 2008.
5. Singhal A. Modern Information Retrieval: A Brief Overview // Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. – 2001. – Vol. 4. – P. 35–43.

6. Tan Z. Fuzzy Metagraph and Its Combination with the Indexing Approach in Rule-Based Systems // IEEE transactions on knowledge and data Engineering. – 2006. – Vol.18, – № 6. – P. 829–841.

7. Thirunavukarasu A., Maheswari S. Fuzzy metagraph based knowledge representation of decision support system // International Journal on Computer Engineering and technology. – 2012. – Vol. 3. – P. 157–166.

8. Thirunavukarasu A., Maheswari S. Technical Analysis of Fuzzy Metagraph based Decision Support System for Capital Market // Journal of Computer Science. – 2013. – Vol. 9. – P. 1146–1155.

References

1. Ivanova G.S. Avtomaticheskij poisk informacii s ispolzovaniem multi-agentnoj sistemy / G.S. Ivanova, A.M. Andreev, V. I. Nefedov, M.A. Shouman, E.V. Egorova. // Jelektromagnitnye volny i jelektronnye sistemy. 2015. T. 20, no. 2. pp. 33–38.
2. Ivanova G.S. Modeli obektov zadach strukturnogo sinteza. // Nauka i obrazovanie. MGTU im. N.Je. Bauman. Jelektron. zhurn. 2006. no. 12. Rezhim dostupa: <http://technomag.edu.ru/doc/62361.html> (data obrashhenija 15.09.2015).
3. Dashore P., Jain S. Fuzzy Metagraph and Hierarchical modeling // International Journal on Computer Science and Engineering. 2011. Vol. 3. no. 1. pp. 435–449.
4. Raghavan P., Schütze H., Manning D. Introduction to Information Retrieval // Cambridge University Press. 2008.
5. Singhal A. Modern Information Retrieval: A Brief Overview // Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 2001. Vol. 4. pp. 35–43.
6. Tan Z. Fuzzy Metagraph and Its Combination with the Indexing Approach in Rule-Based Systems // IEEE transactions on knowledge and data Engineering. 2006. Vol. 18, no. 6. pp. 829–841.
7. Thirunavukarasu A., Maheswari S. Fuzzy metagraph based knowledge representation of decision support system // International Journal on Computer Engineering and technology. 2012. Vol. 3. pp. 157–166.
8. Thirunavukarasu A., Maheswari S. Technical Analysis of Fuzzy Metagraph based Decision Support System for Capital Market // Journal of Computer Science. 2013. Vol. 9. pp. 1146–1155.

Рецензенты:

Черненко В.М., д.т.н., профессор, заведующий кафедрой «Системы обработки информации и управления», Московский государственный технический университет им. Н.Э. Баумана, г. Москва;

Карпенко А.П., д.ф.-м.н., профессор, заведующий кафедрой «Системы автоматизированного проектирования», Московский государственный технический университет им. Н.Э. Баумана, г. Москва.