УДК 025.4:681.3

МОДЕЛЬ ПРОЦЕССА МНОГОЯЗЫКОВОГО ИНТЕЛЛЕКТУАЛЬНОГО ИНФОРМАЦИОННОГО ПОИСКА С УЧЕТОМ МУЛЬТИАГЕНТНОЙ РЕАЛИЗАЦИИ

Шоуман М.А.

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

Рост количества информации в сети Интернет заставляет постоянно совершенствовать средства поиска необходимых данных. С учетом того, что большинство пользователей интернета знают не менее одного иностранного языка, современные средства поиска должны быть многоязыковыми. Однако имеющиеся поисковые системы не обеспечивают автоматического многоязыкового поиска. В статье выполнен анализ процесса многоязыкового интеллектуального информационного поиска и выявлены операции, входящие в указанный процесс. Полученное множество операций декомпозировано для реализации отдельными интеллектуальными агентами. Критериями декомпозиции послужили возможность и целесообразность распараллеливания их выполнения, а также минимизация передаваемой между агентами информации. В результате получена модель многоязыкового интеллектуального информационного поиска в виде нечеткого метаграфа. Предложена реализация поисковой системы на мультиагентной платформе JADA. Приведены результаты выполнения экспериментов на реализованной системе.

Ключевые слова: мультиагентные системы, многоязыковой информационный поиск, нечеткий метаграф

MODEL OF MULTILINGUAL INTELLIGENT INFORMATION SEARCH WITH THE MULTI-AGENT REALIZATION

Shouman M.A.

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

The increasing amount of information on the internet makes us constantly develop the tools that are necessary for data search. Given that most of internet users know at last one foreign language, advanced search tools should be multilingual. Today's search engines don't provide automatic multilingual search.This article gives an analysis of the multilingual intelligent information retrival and and the operation included in this process. The resulting set of informations decomposed to implement individual intelligent agents. Critieria of decomposition are the possibilty of parallelizing their implementation as well as minimizing the transmitted informations between agents. The result is a model of intelligent multilingual information retrival in the form of fuzzy Metagraph. A serach engine using the implementation of the multi-platform JADA is offerd. The results of the experiments with this system are given.

Keywords: multi-agent systems, multilingual information retrieval, fuzzy metagrap.

Многоязыковой информационный поиск – процесс поиска информации в сети Интернет на нескольких языках. При автоматическом многоязыковом поиске пользователь вводит запрос на одном языке, а поисковая система должна обеспечивать перевод запроса на указанные языки и предоставлять ранжированные результаты на всех выбранных языках.

Большинство широко известных поисковых систем, таких как Yandex, Google и др., не реализуют всех функций автоматического многоязыкового поиска. Так, например, поисковые системы Yandex и Google не обеспечивают поиск по переводам введенных терминов. А объявленная возможность многоязыкового поиска Google [3] предполагает поиск по введенному запросу среди документов на других языках. Отсутствие многоязыкового поиска в указанных системах предполагает выполнение перевода терминов и ранжирование результатов поисков на различных языках вручную.

Немногочисленные существующие многоязыковые поисковые системы имеют существенные недостатки, а именно значительное время поиска и сложные методы ранжирования найденных документов, применение которых дополнительно увеличивает время отклика системы.

Уменьшить время многоязыкового поиска можно, используя распараллеливание обработки. С этой целью целесообразным представляется применение мультиагентных платформ для реализации поисковых систем. Помимо обычного распараллеливания мультиагентная организация при наличии соответствующих ресурсов позволяет выполнить масштабирование системы.

Однако для разработки эффективной мультиагентной реализации системы необходимо исследовать процесс многоязыкового информационного поиска и построить его модель.

Целью настоящего исследования является разработка структурной модели

процесса многоязыкового интеллектуального поиска, которая позволит проанализировать процесс поиска и выполнить его декомпозицию для последующей реализации взаимодействующими мультиагентами. Кроме этого, модель процесса должна учитывать неопределенность характеристик документов, на базе которых выполняется оценка релевантности документов запросу.

Модель процесса интеллектуального информационного поиска на трех языках

В [1] рассмотрена модель процесса интеллектуального информационного поиска текстовых документов с помощью мультиагентной системы на одном языке. При этом предполагается выполнение следующих операций:

- ввод запроса;
- поиск в Интернете по ключевым словам;
- извлечение информации из Веб-источников;
- интеллектуальный анализ извлеченных документов;
- получение оценок релевантности документов запросам;
- ранжирование документов по степени релевантности.

Модель представляет собой нечеткий метаграф, поскольку указанная модель является иерархической и позволяет отобразить совокупность операций, выполняемых в процессе поиска одним мультиагентом. Неточность оценок релевантности документов в этой модели отображается в виде нечеткого множества вершин нечеткого метаграфа.

Модель многоязыкового информационного поиска на базе мультиагентной реализации имеет свои особенности. Так, эта модель помимо необходимости перевода терминов, задаваемых пользователем на одном языке, должна учитывать:

- необходимость отдельных поисков в Интернете по запросу на каждом языке;
- особенности интеллектуального анализа текстов на различных языках;
- необходимость совместного ранжирования полученных результатов.

Первая операция многоязыкового поиска – ввод запроса. Эту операцию целесообразно выполнять отдельным интерфейсным агентом.

Следующая операция – перевод на указанные языки. Эту операцию должен выполнять агент-переводчик. Анализ показал, что при переводе по возможности следует использовать словарь терминов (устоявшихся словосочетаний), а не отдельных слов [6, 9]. Это обеспечит лучшее качество перевода и исключит извлечение нерелевантных документов.

Операции поиска выполняются отдельно с разными переводами запроса. Однако за счет индексирования и других приемов ускорения время поиска по сравнению с временем выполнения других операций невелико, а потому поиск целесообразно выполнять одним агентом, обращаясь для получения ссылок к одной из поисковых систем (Google, Yandex и т.п.).

Операции извлечения и анализа текстов выполняются для каждого языка отдельно, поскольку документы на различных языках будут обрабатываться с использованием соответствующих правил языков. Интеллектуальный анализ текстов включает лексемизацию, фильтрацию и лемматизацию, что позволяет увеличить точность оценки релевантности документов [7, 8].

В [2] для оценки релевантности документов предложено использовать векторную меру и статистические веса появления терминов в документах. Неточность этих оценок частично скомпенсирована за счет использования экспертной системы на базе нечеткого логического вывода Сугено [2]. Правила для системы нечеткого логического вывода формируются с использованием характеристик вершин нечеткого метаграфа.

Операция оценки релевантности уже не зависит от языка, поскольку на этом этапе для всех документов должна использоваться единая мера сходства. Однако выполнение этой операции разными агентами для различных языков увеличит возможности распараллеливания процесса.

Операции извлечения, анализа текстов и оценки релевантности передают друг другу большие объемы информации (тексты документов). Передача этих данных различными агентами через сообщения нецелесообразна, поэтому эти операции следует выполнять одним агентом для каждого языка.

Операция ранжирования должны выполняться над всеми документами сразу, поэтому ее также должен выполнять один агент.

Тогда структура нечеткого метаграфа $\tilde{S}$, представляющего собой модель процесса многоязыкового интеллектуального поиска, описывается следующим образом:

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

где $X = \left\{ x_i, i = \overline{1,23} \right\}$ – множество операций, осуществляемых в процессе поиска и интеллектуальной обработки документов; $\tilde{X}$ – нечеткое множество на X – множество операций (табл. 1), осуществляемых в процессе поиска и интеллектуальной обработки документов с учетом неопределенности

$\tilde{x}_i = (x_i, \mu(x_i))$, где μ – функции принадлежности, $\forall x_i \in X \cdot \mu: x_i \rightarrow [0, 1]$;

$\tilde{E} = \{\tilde{e}_2, \tilde{e}_2, \tilde{e}_3, \tilde{e}_4, \tilde{e}_5, \tilde{e}_6, \tilde{e}_7, \tilde{e}_8\}$ – множество сообщений, передаваемых между мультиагентами:

$\tilde{e}_1 = \langle \{\tilde{x}_1\}, \{\tilde{x}_2, \tilde{x}_3, \tilde{x}_4\} \rangle$ – передается запрос;

$\tilde{e}_2 = \langle \{\tilde{x}_2, \tilde{x}_3, \tilde{x}_4\}, \{\tilde{x}_5, \tilde{x}_6, \tilde{x}_7\} \rangle$ – передаются запрос и его переводы;

$\tilde{e}_3 = \langle \{\tilde{x}_5, \tilde{x}_6, \tilde{x}_7\}, \{\tilde{x}_8, \tilde{x}_9, \tilde{x}_{10}, \tilde{x}_{11}, \tilde{x}_{12}\} \rangle$ – передаются извлеченные тексты документов;

$\tilde{e}_4 = \langle \{\tilde{x}_5, \tilde{x}_6, \tilde{x}_7\}, \{\tilde{x}_{13}, \tilde{x}_{14}, \tilde{x}_{15}, \tilde{x}_{16}, \tilde{x}_{17}\} \rangle$ – передаются извлеченные тексты документов;

$\tilde{e}_5 = \langle \{\tilde{x}_5, \tilde{x}_6, \tilde{x}_7\}, \{\tilde{x}_{18}, \tilde{x}_{19}, \tilde{x}_{20}, \tilde{x}_{21}, \tilde{x}_{22}\} \rangle$ – передаются извлеченные тексты документов;

$\tilde{e}_6 = \langle \{\tilde{x}_8, \tilde{x}_9, \tilde{x}_{10}, \tilde{x}_{11}, \tilde{x}_{12}\}, \{\tilde{x}_{23}\} \rangle$ – передаются оценки релевантности документов;

$\tilde{e}_7 = \langle \{\tilde{x}_{13}, \tilde{x}_{14}, \tilde{x}_{15}, \tilde{x}_{16}, \tilde{x}_{17}\}, \{\tilde{x}_{23}\} \rangle$ – передаются оценки релевантности документов;

$\tilde{e}_8 = \langle \{\tilde{x}_{18}, \tilde{x}_{19}, \tilde{x}_{20}, \tilde{x}_{21}, \tilde{x}_{22}\}, \{\tilde{x}_{23}\} \rangle$ – передаются оценки релевантности документов.

Таблица 1

Условные обозначения элементарных операций поиска

Элемент множества	Обозначение	Операция
$\tilde{x}_1$	UI	Ввод ключевых слов пользователем
$\tilde{x}_{2,3,4}$	AR, RU, EN	Перевод ключевых слов на остальные языки
$\tilde{x}_{5,6,7}$	URL	Поиск ссылок на каждом языке
$\tilde{x}_{8,13,18}$	RD	Извлечение документов
$\tilde{x}_{9,14,19}$	DT	Лексемизация документов
$\tilde{x}_{10,15,20}$	DF	Фильтрация документов
$\tilde{x}_{11,16,21}$	DS	Лемматизация документов
$\tilde{x}_{12,17,22}$	TW	Вычисление весов терминов
$\tilde{x}_{23}$	OE	Вычисление оценок релевантности и ранжирование результатов

Каждому агенту в модели соответствует подмножество операций:

- интерфейсному агенту – $\tilde{X}_1 = \{\tilde{x}_1\}$;
- агенту-переводчику – $\tilde{X}_2 = \{\tilde{x}_2, \tilde{x}_3, \tilde{x}_4\}$;
- поисковому агенту – $\tilde{X}_3 = \{\tilde{x}_5, \tilde{x}_6, \tilde{x}_7\}$;
- агенту извлечения документов и интеллектуальной обработки текстов на арабском языке – $\tilde{X}_4 = \{\tilde{x}_8, \tilde{x}_9, \tilde{x}_{10}, \tilde{x}_{11}, \tilde{x}_{12}\}$;
- агенту извлечения документов и интеллектуальной обработки текстов на русском языке – $\tilde{X}_5 = \{\tilde{x}_{13}, \tilde{x}_{14}, \tilde{x}_{15}, \tilde{x}_{16}, \tilde{x}_{17}\}$;
- агенту извлечения документов и интеллектуальной обработки текстов на английском языке – $\tilde{X}_5 = \{\tilde{x}_{18}, \tilde{x}_{19}, \tilde{x}_{20}, \tilde{x}_{21}, \tilde{x}_{22}\}$;
- агенту ранжирования результатов поиска – $\tilde{X}_6 = \{\tilde{x}_{23}\}$.

Аналогично могут быть построены модели информационного поиска, рассчитанные на большее количество языков. Для добавления еще одного языка следует добавить словари для перевода терминов на этот язык и интеллектуального агента, умеющего выполнять анализ текстов на этом языке.

Реализация системы многоязыкового интеллектуального поиска на мультиагентной платформы JADA

Процесс многоязыкового информационного поиска был реализован в виде мультиагентной системы. В качестве программной среды при этом была использована среда JADE (Java Agent Development Framework) [5]. Эта среда является программным обеспечением для разработки агентских приложений в соответствии со спецификациями FIPA [4] для взаимозаменяемых интеллектуальных мультиагентных систем.

На рис. 2 показан вывод агента-переводчика пользовательского запроса «database».

Выполненная реализация позволила исследовать процесс интеллектуального многоязыкового информационного поиска. Так, в табл. 2 показаны результаты вычисления весов терминов для запроса «Искусственный интеллект». Указанный запрос был переведен на английский и арабский языки, и полученные три запроса были переданы поисковой системе Google. Анализ представленных результатов подтверждает необходимость совокупного ранжирования полученных документов, поскольку среди первых 7 найденных документов на каждом языке присутствуют документы с низкими характеристиками весов терминов, т.е. документы, которые на основании вычисленных характеристик релевантности должны быть признаны нерелевантными.

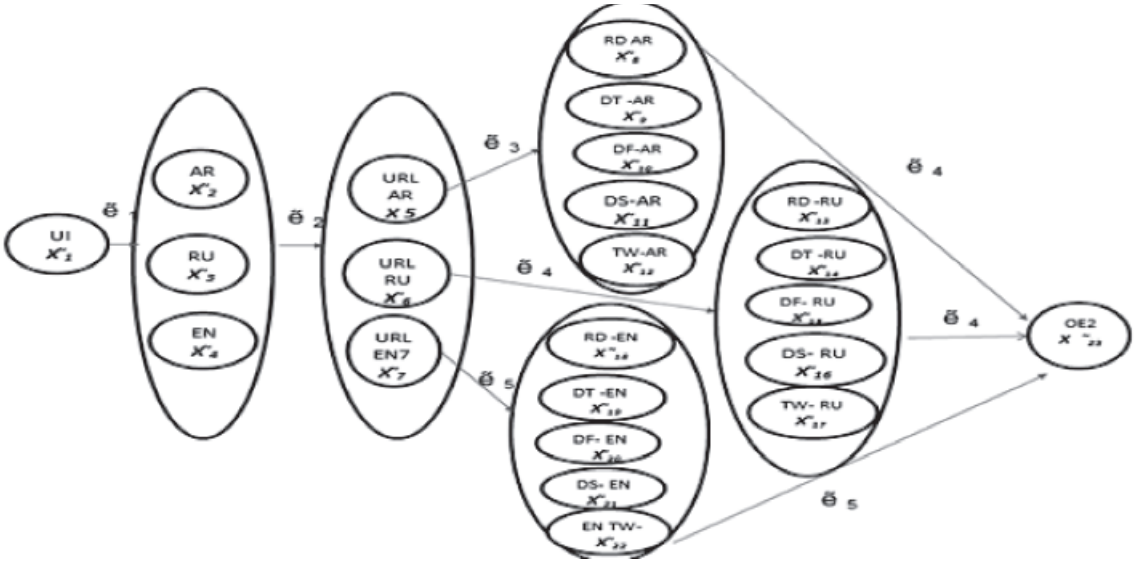

Рис. 1. Модель процесса интеллектуального информационного поиска на трех языках: RU – русском; EN – английском; AR – арабском

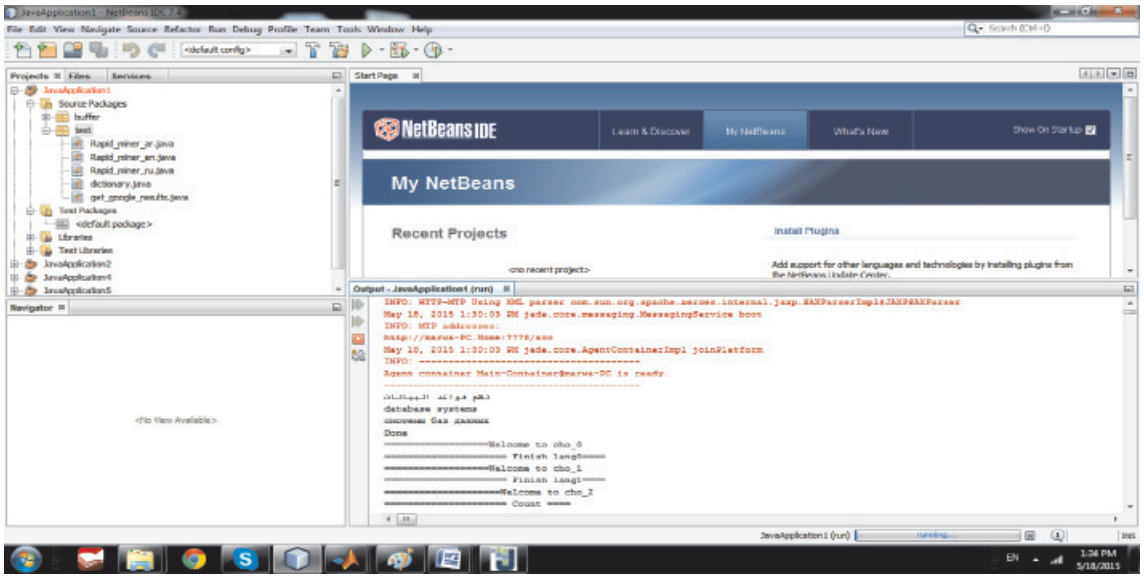

Рис. 2. Вывод агента-переводчика пользовательского запроса со словом «database»

Таблица 2

Взвешенные нормализованные веса терминов для первых семи документов, найденных по запросу «Интеллект искусственный»

Арабский язык. Нормализованный вес термина		Английский язык. Нормализованный вес термина		Русский язык. Нормализованный вес термина	
كذي	صانطعاي	Artifici	Intellig	Интеллект	искусственный
0,812	0,583	0,456	0,494	0,781	0,625
0	0	0,832	0,555	0,707	0,707
0,657	0,478	0,024	0,975	0,707	0,707
0	0	0,618	0,786	0,708	0,707
0,994	0,110	0	0	0,723	0,690
0	0	0,678	0,735	0,707	0,707
0,942	0,099	0,707	0,707	0,669	0,706

На рис. 3 показана диаграмма прироста времени поиска при добавлении одного языка. За счет распараллеливания обработки этот прирост составляет примерно 25 ± 10 %, что зависит от количества документов, найденных в Интернете по конкретному запросу на добавленном языке.

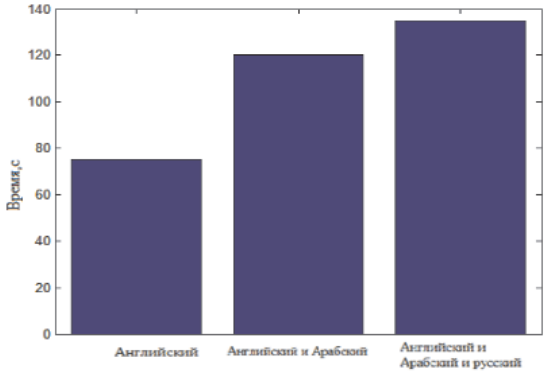

Рис. 3. Зависимость времени работы системы от количества используемых языков

Эксперимент был осуществлен с использованием следующего оборудования: ЦП Intel (R) core (TM) i7-2670 QM CPU @2.20 GHZ, установленная память (ОЗУ): 6.00 ГБ и тип ОС: 64-битная операционная система.

Заключение

В статье предложена модель многоязыкового интеллектуального информационного поиска в виде нечеткого метаграфа, которая позволяет отобразить основные особенности указанного процесса с учетом последующей реализации поисковой системы на мультиагентной платформе.

Исследование выполненной реализации позволяет сделать вывод об обязательности ранжирования совокупности результатов выполнения многоязыковых поисковых запросов.

Статистическая обработка результатов исследования мультиагентной реализации позволяет сделать вывод о том, что прирост времени при добавлении одного языка зависит от количества документов на этом языке, найденных при выполнении запросов, и составляет примерно 25 ± 10 %.

Список литературы

1. Иванова Г.С. Автоматический поиск информации с использованием мульти-агентной системы / Г.С. Иванова, А.М. Андреев, В.И. Нефедов, М.А. Шоуман, Е.В. Егорова // Электромагнитные волны и электронные системы. – 2015. – Т. 20 – № 2. – С. 33–38.

2. Иванова Г.С., Андреев А.М., Шоуман М.А. Поиск и Ранжирование документов с использованием мультиагентной системы. //Фундаментальные исследования. – 2015. – № 10 . – Часть 3. – С.489–494.

3. Многоязычный поиск Google. – URL: http://www.searchengines.ru/seoblog/ mnogoyazych.html (дата обращения 11.11.2015).

4. Bellifemine, F. L., Caire, G., и др. Developing Multi-Agent Systems with JADE. : Wiley. – 2007.

5. Bellifemine F, Poggi A, Rimassa G: JADE – A FIPA-compliant agent framework, University of Parma. – 2000.

6. Jialun Q., Zhou Y., Yilu Z., Chau M., Hsinchun C. Multilingual Web retrieval: An experiment in English-Chinese business intelligence. // Journal of the American Society for Information Science and Technology(JASIST). – 2006. – Vol. 5. –P. 671–683.

7. Manning D., Raghavan C., Schütze H. Introduction to Information Retrieval: Cambridge. – England. – 2008.

8. Singhal A. Modern information retrieval: a brief overview // Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. – 2001. – Vol 24. – P. 35–43.

9. Yoshinaga K., Terano T., Zhong N. Multi-lingual Intelligent Information Retriever with Automated Ontology Generator // In Proc. of Third International Conference on Knowledge-Based Intelligent Information Engineering Systems. – 1999. – P. 62–65.

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., Andreev A.M., Shouman M.A. Poisk i Ranzhirovanie dokumentov s ispolzovaniem multiagentnoj sistemy // Fundamentalnye issledovanija. 2015. no 10. pp. 489–494.

3. Mnogojazychnyj poisk Google. URL: http://www.searchengines.ru/seoblog/ mnogoyazych.html (accessed 11.11.2015).

4. Bellifemine F.L., Caire G. i dr. Developing Multi-Agent Systems with JADE: Wiley, 2007.

5. Bellifemine F., Poggi A, Rimassa G: JADE – A FIPA-compliant agent framework, University of Parma. – 2000.

6. Jialun Q., Zhou Y., Yilu Z., Chau M., Hsinchun C. Multilingual Web retrieval: An experiment in English-Chinese business intelligence // Journal of the American Society for Information Science and Technology(JASIST). 2006. Vol. 5. pp. 671–683.

7. Manning D., Raghavan C., Schütze H. Introduction to Information Retrieval: Cambridge. England. 2008.

8. Singhal A. Modern information retrieval: a brief overview // Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 2001. Vol. 24. pp. 35–43.

9. Yoshinaga K., Terano T., Zhong N. Multi-lingual Intelligent Information Retriever with Automated Ontology Generator // In Proc. of Third International Conference on Knowledge-Based Intelligent Information Engineering Systems. 1999. pp. 62–65.