МЕТОДЫ АВТОМАТИЗИРОВАННОГО ВОПРОСНО-ОТВЕТНОГО ПОИСКА И ИХ КОМПЛЕКСНОЕ ИСПОЛЬЗОВАНИЕ

METHODS OF AUTOMATED QUESTION-ANSWER SEARCH AND THEIR COMPLEX USE

V. Simankov
D. Tolkachev

Summary. The methods used in the question-answer search are analyzed and classified. The method of complex use of methods for achieving the greatest flexibility and efficiency in the search for answers to the question is proposed. The results of comparative analysis of searching for answers are presented for different systems.

Keywords: answers, question, knowledge base, automatic summarization, ternary expressions.

Симанков Владимир Сергеевич
Д.т.н., профессор, ФГБОУ ВО «Кубанский государственный технологический университет», vs@simankov.ru
Толкачев Демид Максимович
К.т.н., старший преподаватель, ФГБОУ ВО «Кубанский государственный технологический университет», Gendalf373@rambler.ru

Аннотация. Проанализированы и классифицированы методы, применяемые в вопросно-ответном поиске. Предложена методика комплексного использования методов для достижения наибольшей гибкости и эффективности при поиске ответов на вопрос. Представлены результаты сравнительного анализа поиска ответов различными системами.

Ключевые слова: ответ; вопрос; база знаний; автореферирование; тернарные выражения.

Вопросно-ответный поиск (ВОП) представляет собой особый тип информационного поиска. Учитывая набор документов, система вопросно-ответного поиска (СВОП) пытается найти правильный ответ на вопрос, заданный на естественном языке. ВОП включает в себя информационные технологии, искусственный интеллект, обработку естественного языка, управление базами данных и знаний, когнитивные технологии и технологии автореферирования.

Все СВОП можно разделить на две большие категории:

1) системы, обрабатывающие тексты на естественном языке;

2) системы, работающие с базой знаний (БЗ).

Первая категория СВОП использует в качестве источника ответов произвольные тексты. Они могут быть как загружены в систему предварительно, так и выбираться в процессе её работы, например, из сети Интернет. Для СВОП этой категории важным аспектом являются методы и алгоритмы анализа текстов на естественном языке.

Вторая категория СВОП использует в качестве источника ответов базу знаний. Такая БЗ может как создаваться вручную экспертами, так и автоматизированно, с привлечением специалистов лишь для проверки занесённых в неё фактов. Для СВОП этой категории важным аспектом являются методы наполнения БЗ и работы с ней.

Существуют и другие классификации СВОП, например, приведённая в [1]:

- веб-ориентированные СВОП;
- СВОП, основанные на поиске и извлечении информации;
- узкоспециализированные СВОП;
- СВОП на основе правил.

Существуют различные методы, используемые в автоматизированном вопросно-ответном поиске. Их можно разделить на основные и вспомогательные. Основные методы используются для непосредственного нахождения ответа на вопрос, тогда как вспомогательные применяются для облегчения этой задачи, повышения эффективности или быстродействия, но не способны самостоятельно найти ответ.

В [2] рассматривались такие методы и алгоритмы вопросно-ответного поиска:

- Ключевые слова;
- Семантический анализатор В. Тузова;
- Тернарные выражения;
- Шаблоны;
- Активные семантические сети Поспелова;
- Алгоритм N-грамм;
- Алгоритм синтаксического внутрисегментного анализа текста.

Методы, рассмотренные в [2], можно расширить нейронными сетями [3], [4], [5], [6]. Также следует заметить, что Семантический анализатор В. Тузова является одной из вариаций семантического анализа, Активные семантические сети Поспелова — одной из вариаций семантических сетей, а Алгоритм синтаксического вну-

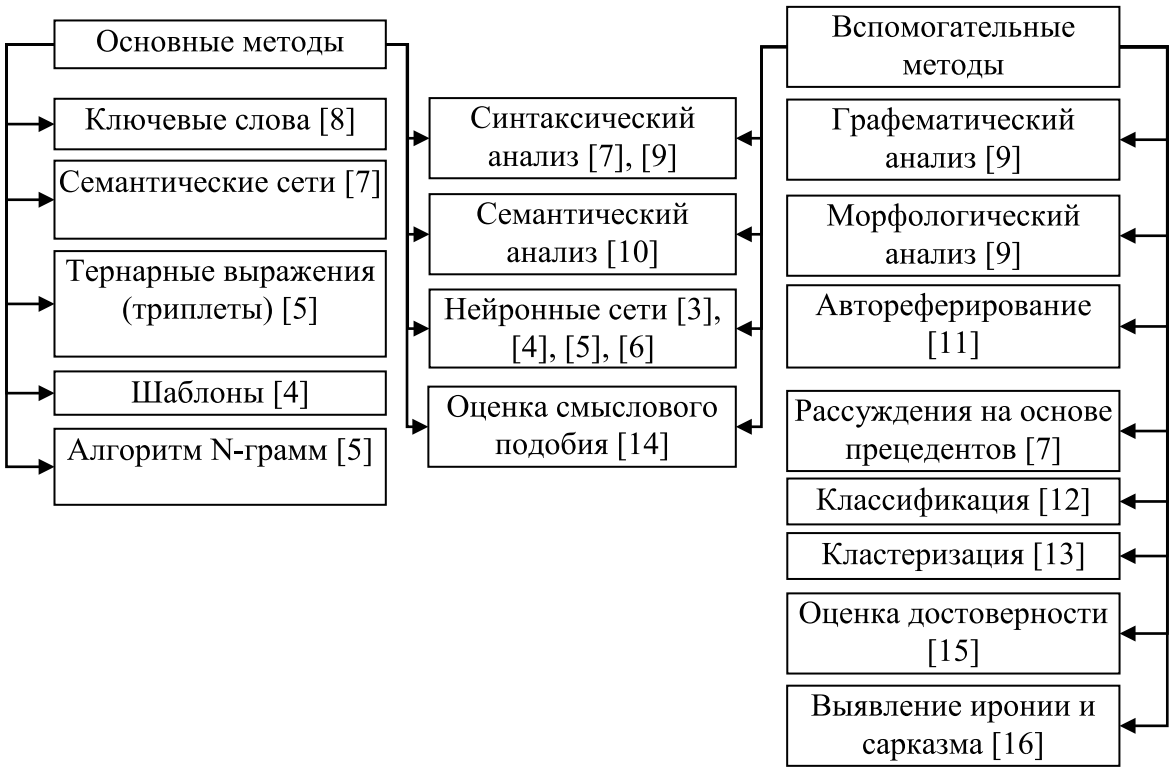

Рис. 1. Схема методов автоматизированного вопросно-ответного поиска

трисегментного анализа текста — одной из вариаций синтаксического анализа.

На основе [2] и с учётом актуальных работ в данной области систематизируем перечень методов и алгоритмов вопросно-ответного поиска (рисунок 1).

На схеме приведены все основные методы, используемые в вопросно-ответном поиске. Комбинируя их, можно построить любую существующую систему поиска ответов на вопрос. Эффективность каждого метода во многом определяется его конкретной модификацией и особенностями её реализации на практике, поэтому невозможно однозначно выявить лучший метод. Однако можно утверждать, что ни один из методов в отдельности не обеспечивает эффективного механизма автоматизированного поиска ответов на вопросы, поэтому необходимо их комплексное использование. Предложим следующую методику комплексного использования части рассмотренных методов.

Система вопросно-ответного поиска будет содержать две основные части: модуль работы с базой знаний и модуль работы с сетью Интернет. Модуль работы с базой знаний будет осуществлять её пополнение и актуализацию, а также поиск в ней фактов и выдачу их в качестве ответов пользователям.

Модуль работы с сетью Интернет будет осуществлять взаимодействие с поисковыми системами и анализ полученных веб-источников с целью выявления ответов на вопросы.

Система будет обрабатывать пользовательский вопрос в несколько этапов:

- анализ вопроса;
- обращение к базе знаний и поиск в ней ответа на вопрос;
- обращение к сети Интернет и поиск в ней ответов в случае, если в БЗ отсутствует однозначный ответ;
- актуализация БЗ на основе обратной связи с пользователем.

Предложенная методика развивает идеи, рассмотренные в [2], и является в достаточной степени универсальной. Она позволяет как проводить поиск ответов по сформированной базе знаний, так и искать новые сведения в сети Интернет, актуализируя БЗ по мере необходимости. Схематично изобразим её (рисунок 2).

Когда ответ на вопрос отсутствует в БЗ, осуществляется анализ веб-источников с целью его поиска. Для этого направляется запрос к одной из современных поисковых систем (Google, Яндекс и т.д.), и полученные

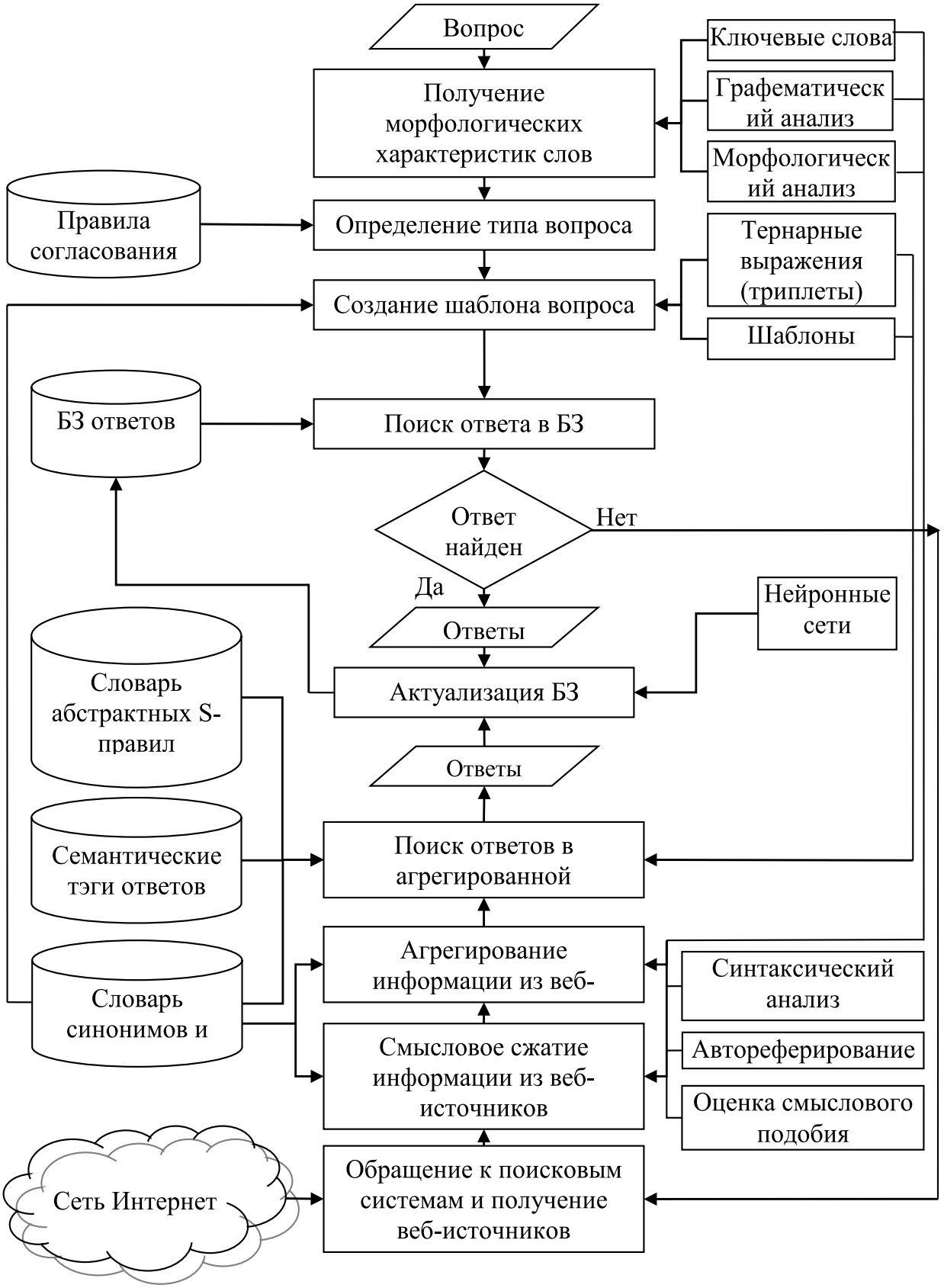

Рис. 2. Схема предложенной методики вопросно-ответного поиска

Таблица 1. Результаты поиска ответов различными системами

Критерий	AskNet	START	IntellST
Общее число релевантных ответов	15	5	36
Общее число данных ответов	26	6	53
Число вопросов, на которых был дан правильный и прямой ответ	8	5	23
Число релевантных ответов, допускающих исключение	6	0	9
Точность	57,69%	83,33%	67,92%
Полнота	32,00%	20,00%	92,00%
Избыточность	40,00%	0,00%	25,00%
Сбалансированная F-мера	41,17%	32,26%	78,15%

веб-страницы анализируются с использованием методики проблемно-ориентированного автореферирования (ПОА) [17].

Методика ПОА предполагает анализ html-кода веб-страниц: он разбивается на абзацы, они очищаются от html-тэгов и делятся на предложения. Затем осуществляется определение наиболее значимых предложений веб-страниц на основе наборов индикаторов, построенных как с использованием анализа вопроса, так и специальной базы знаний. Также осуществляется выявление и исключение предложений, которые не могут служить ответом на вопрос. Это слишком короткие предложения, а также предложения, полученные из элементов меню веб-страницы.

Чтобы предложения автореферата не выглядели вырванными из контекста, нужно определять семантические связи между ними. Выявление семантических связей в ПОА осуществляется с помощью набора построенных с учётом синтаксиса и семантики языка правил в виде регулярных выражений. Предложения с установленной сильной сематической связью не будут разделяться при формировании автореферата.

При автореферировании также необходимо определять смысловое подобие фраз для исключения дублирующих друг друга. С учётом признаков смыслового сходства была предложена усовершенствованная методика расчёта смыслового подобия предложений [17].

Из отдельных авторефератов проанализированных веб-источников формируется общий, интегрированный автореферат. Для его формирования также была предложена своя методика [17].

Поиск конкретных ответов на вопрос осуществляется уже в рамках интегрированного автореферата. Для этого используется специализированная методика, в которой применяются тернарные выражения, шаблоны, а также результаты предшествующего анализа источников [18], [19].

На основании предложенной методики была реализована система вопросно-ответного поиска IntellST [20]. Сравнительный анализ её эффективности с отечественной русскоязычной системой AskNet [21] и зарубежной англоязычной системой START [22] по поиску ответов на 25 вопросов приведён в таблице 1 [2].

Результаты системы IntellST в таблице 1 были получены с использованием исключительно поиска ответов в сети Интернет. Как видно из результатов, IntellST превосходит свои аналоги в решении задачи вопросно-ответного поиска.

Таким образом, вопросно-ответный поиск интенсивно развивается в настоящее время. Исследователи больше внимания уделяют СВОП, работающим с базами знаний, поскольку эта категория вопросно-ответных систем позволяет достичь большей точности и достоверности. Однако в БЗ не может содержаться вся накопленная человечеством информация. Также существует проблема актуализации знаний. Поэтому имеет смысл создание гибридных систем, совмещающих базу знаний с анализом интернет-источников, как наиболее полных и актуальных.

В статье проанализированы и классифицированы методы, применяемые в вопросно-ответном поиске. Предложена методика комплексного использования методов для достижения наибольшей гибкости и эффективности при поиске ответов на вопрос. Создана система вопросно-ответного поиска IntellST, анализирующая интернет-источники и предполагающая возможность использования базы знаний. Проведённый сравнительный анализ эффективности системы показал её практическую применимость в задаче вопросно-ответного поиска.

ЛИТЕРАТУРА

1. Poonam Gupta, Vishal Gupta. A Survey of Text Question Answering Techniques. International Journal of Computer Applications (0975–8887), Volume 53– No.4, September 2012.— 8 p.
2. Владимир Симанков, Демид Толкачев. Поиск информации в Интернете. Подходы, методы и алгоритмы. LAP LAMBERT Academic Publishing, 2016.— 296 с. ISBN-13: 978–3–659–90123–2.
3. Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daume III. A Neural Network for Factoid Question Answering over Paragraphs. Empirical Methods in Natural Language Processing, 2014, 12 pages.
4. Wen-tau Yih, Ming-Wei Chang, Xiaodong He, Jianfeng Gao. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the AFNLP, July 28, 2015.
5. Antoine Bordes, Nicolas Usunier, Sumit Chopra, Jason Weston. Large-scale Simple Question Answering with Memory Networks. arXiv:1506.02075v1 [cs.LG] 5 Jun 2015, 10 p.
6. Caiming Xiong, Victor Zhong, Richard Socher. Dynamic Coattention Networks For Question Answering. arXiv:1611.01604v2 [cs.CL] 17 Nov 2016.— 13 p.
7. Karl-Heinz Weis. A Case Based Reasoning Approach for Answer Reranking in Question Answering. In Proceedings Informatik 2013, Koblenz, Germany, 2013.— 12 p. arXiv:1503.02917v1.
8. Wen-tau Yih, Ming-Wei Chang, Christopher Meek, Andrzej Pastusiak. Question Answering Using Enhanced Lexical Semantic Model. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1744–1753, Sofia, Bulgaria, August 4–92013.
9. Ким К. Х., А. П. Савинов. Синтаксический анализатор для вопросно-ответной системы. Известия Томского политехнического университета,— Т. 315.— № 5,— 2009.— с. 133–138.
10. Мозговой Максим Владимирович. Машинный семантический анализ русского языка и его применения. Диссертация на соискание ученой степени кандидата физико-математических наук. Санкт-Петербург, 2006.— 116 с.
11. Хорошевский В. Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1) / В. Ф. Хорошевский // Искусственный интеллект и принятие решений 1/2008.— с. 80–97.
12. Задача классификации [Электронный ресурс]. Режим доступа: http://www.intuit.ru/studies/courses/6/6/lecture/166?page=1#sect2 (12.05.2017).
13. Задача кластеризации [Электронный ресурс]. Режим доступа: http://www.intuit.ru/studies/courses/6/6/lecture/166?page=4#sect7 (12.05.2017).
14. Захаров В. Н. Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний / В. Н. Захаров, А. А. Хорошилов // Труды XIV-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции».— RCDL'2012, Переславль-Залесский, Россия, 15–18 октября 2012 г.
15. Марманис Х., Бабенко Д. Алгоритмы интеллектуального Интернета. Передовые методики сбора, анализа и обработки данных.— Пер. с англ.— СПб.: Символ-Плюс, 2011.— 480 с., ил. ISBN978–5–93286–186–8.
16. Bharti S. K., Babu K. S., Jena S. K. Parsing-based Sarcasm Sentiment Recognition in Twitter Data //Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015.— ACM, 2015.— С. 1373–1380.
17. Симанков В. С. Автореферирование с определением смысловой связности и использованием мер включения для поиска ответов на вопросы в сети Интернет / В. С. Симанков, Д. М. Толкачев // Наука и бизнес: пути развития, № 12 (66), М., 2016.— с. 30–34.
18. Симанков В. С. Методические положения автоматического поиска ответов на вопросы / В. С. Симанков, Д. М. Толкачев // Перспективы науки, № 9 (60), Тамбов, 2014.— с. 80–85.
19. Симанков В. С. Разработка информационно-аналитической системы получения релевантных данных и знаний в сети Интернет / В. С. Симанков, Д. М. Толкачев // Программные системы и вычислительные методы.— 2014.— № 3.— С. 298–311. DOI: 10.7256/2305–6061.2014.3.13396.
20. Интеллектуальная информационно-аналитическая система поиска ответов в сети Интернет IntellST / В. С. Симанков, Д. М. Толкачев; — № 2015619195; заявка № 2015615783 от 30.06.2015; зарегистрировано в реестре программ для ЭВМ 26.08.2015.
21. Семантическая поисковая система AskNet [Электронный ресурс]. Режим доступа: http://www.asknet.ru/ (12.05.2017).
22. START, Natural Language Question Answering System [Электронный ресурс]. Режим доступа: http://start.csail.mit.edu/index.php (12.05.2017).

Журнал «Современная наука: актуальные проблемы теории и практики»