УДК 002.53:004.89

05.00.00 Технические науки

СЕМАНТИЧЕСКАЯ ПОИСКОВАЯ СИСТЕМА НА ОСНОВЕ ОТОБРАЖЕНИЯ ОНТОЛОГИИ

Новикова Юлия Сергеевна
аспирант каф. САПР

Терещенко Дмитрий Юрьевич
аспирант каф. САПР

Поляков Сергей Евгеньевич
магистрант каф. Вычислительной техники

Самойлов Алексей Николаевич
к.т.н., зав.каф. Вычислительной техники
e-mail: asamoylov@sfedu.ru
Институт компьютерных технологий и информационной безопасности Южного федерального университета, Россия

Мы живем в мире быстро развивающихся информационных технологий, где многие организации связаны друг с другом, формируя сложные информационные сети. Поэтому хранение, анализ и извлечение информации является очень сложной и актуальной задачей. В связи с появлением поисковых систем и средств связи миллионы людей занимаются поиском и извлечением информации. Коммерческие поисковые системы, такие как Google, используют поиск ключевых слов, основанный на логических запросах. Основным недостатком такого поиска является то, что он возвращает много нерелевантной информации, что приводит к низкой точности. В данной статье мы фокусируемся на построении системы семантического поиска, основанной на отображении онтологической модели. Это включает в себя различные этапы разработки онтологии, фазы индексирования и поиска информации

Ключевые слова:ОНТОЛОГИЯ, СЕМАНТИЧЕСКАЯ СЕТЬ, ПОИСК ИНФОРМАЦИИ, ОТОБРАЖЕНИЕ ОНТОЛОГИИ

Doi: 10.21515/1990-4665-133-014

UDC 002.53:004.89

Technical sciences

SEMANTIC SEARCH ENGINE BASED ON ONTOLOGY MAPPING

Novikova Yulia Sergeevna
Postgraduate student of CAD department

Tereshenko Dmitriy Yurievich
Postgraduate student of CAD department

Polyakov Sergey Evgenievich
graduate student of the Department of Computer Science

Samoylov Alexey Nikolaevich
Cand.Tech.Sci., Head of the Department of Computer Science,
e-mail: asamoylov@sfedu.ru
Institute of Computer Technology and Information Security, Southern Federal University, Russia

We live in a world of rapidly developing information technologies, where many organizations are related to each other, forming complex information networks. Therefore, storage, analysis and retrieval of information is a very complex and urgent task. In connection with the advent of search engines and communications, millions of people are engaged in the search and retrieval of information. Commercial search engines, such as Google, use keyword search based on logical queries. The main disadvantage of this search is that it returns a lot of irrelevant information, which leads to low accuracy. In this article, we focus on building a semantic search system based on the mapping of the ontological model. This includes various stages of development of ontology, the phase of indexing and information retrieval

Keywords: ONTOLOGY, SEMANTIC WEB, INFORMATION RETRIEVAL, ONTOLOGY MAPPING

1. ВВЕДЕНИЕ

В эпоху цифровых технологий все документы связаны друг с другом через различные отношения, образующие сложную информационную сеть.

В интернете мы сталкиваемся с различными проблемами при хранении и извлечении информации. Поэтому мы полагаемся на поисковые системы для преодоления этих трудностей. Стремясь преодолеть ограничения моделей на основе ключевых слов, идея семантического поиска, понимаемая как поиск по смыслу, а не по литералам, была в центре исследований в области поиска информации и сообщества *Semantic Web* [1].

Семантический поиск был представлен в области информационного поиска с начала 80-х годов. Большинство подходов к поиску информации применяют лингвистические алгоритмы, основанные на структурах обработки данных и таксономиях человеческого языка, где уровень концептуализации часто невелик и разрежен, особенно уровень отношений, которые обычно лежат в основе выражения пользовательских потребностей и поиска ответов. Онтология применяется в поиске информации, где используются базы знаний, усиливающие семантический поиск, с одной стороныуправляя использованием полноценных семантических онтологий, а с другой – рассматривает неструктурированный контент как целевое пространство поиска. Другими словами, в этой работе исследуется использование семантической информации для поддержки более сложных запросов и более точных результатов,в то время как проблема поиска формулируется таким образом, который согласуется с информационным поисковым пространством, что обеспечивает более точные и применимые подходы [2].

2. SEMANTICWEB

Текущая всемирная паутина (*WWW*) – это огромная библиотека взаимосвязанных документов, которые передаются спомощью компьютеров и предоставляются людям.

SemanticWeb – это попытка расширить существующую сеть, чтобы компьютеры могли обрабатывать информацию, представленную на *WWW*, интерпретировать и связывать ее, чтобы помочь людям найти необходимые знания. Подобно тому, как *WWW* является огромной распределенной гипертекстовой системой, семантическая сеть предназначена для формирования огромной распределенной системы на основе знаний. В центре семантической сети – обмен данными, а не документами. Другими словами, это проект, который должен обеспечить общую структуру, которая позволяет делиться данными и повторно использовать их в рамках приложений, предприятия и сообщества. Это совместная работа под руководством консорциума *World Wide Web* (*W3C*). Семантическая сеть достигла своего пика в последние годы благодаря использованию явного представления метаданных информации в сети. Представление метаданных встроено в веб-страницу с использованием *RDF* для улучшения визуализации результатов [3].

SemanticWeb обычно строится на синтаксисах, которые используют *URI* идентификаторы для представления данных, как правило, в структурах, основанных на триплетах, их также называют синтаксисами «структурой описания ресурсов» [4-7, 9].

3. ТИПЫ СЕМАНТИЧЕСКОГО ПОИСКА

Классификация подходов семантического поиска сложна не только из-за их описаний в литературе, но и из-за большого числа измерений, участвующих в задаче поиска информации.В этом разделе предлагается набор общих критериев, по которым можно классифицировать и сравнивать *SemanticWeb* и подходы к поиску информации, определяя их основные преимущества и ограничения.

Представление семантических знаний. В литературе можно выделить три основные тенденции, основанные на типе и использовании представления семантических знаний [6, 9]:

1. статистические подходы – используют статистические модели для идентификации групп слов, которые обычно появляются вместе, и поэтому могут совместно описывать конкретное понятие;

2. подходы лингвистической концептуализации основаны на легкой концептуализации, обычно рассматривающей несколько типов отношений между концепциями с низкими уровнями указания информации;

3. онтологические подходы, которые рассматривают предметные области, представленные в форме основанных на онтологиях баз знаний.

Важный аспект, который характеризует семантические модели поиска – это способ, которым пользователь выражает свои информационные потребности. Подходы семантического поиска могут быть охарактеризованы тем, нацелены ли они на получение данных или информации, в то время как большинство подходов информационного поиска (*IR*) возвращает документы в ответ на запросы пользователей.

Семантическое сходство. Модель векторного пространства (*VSM*) – используется для представления документов с помощью слов, которые они содержат. Как правило, для семантического поиска информации используется *VSM* и набор документов, а традиционная модель поиска информации обычно измеряет сходство запроса и разных документов, а затем возвращает документы с наивысшим сходством в качестве результатов. Как и в модели векторного пространства, запрос и документы рассматриваются как некоторая точка на векторном пространстве. Подобное сходство между двумя документами воспринимается косинусом угла между векторами. Чем меньше значение угла косинуса, тем больше сходство.

Мы разрабатываем центральную онтологию, которая используется каждым модулем системы, особенно на этапах вывода и извлечения информации. Таким образом, общая производительность системы сильно зависит от качества разработанной онтологии. Мы следуем итеративному

процессу разработки на этапе формирования онтологии. Сначала мы начали с базовой онтологии, включающейосновные концепции и единую иерархию. Затем мы экспериментировали с этой онтологией и исправляли проблемы в рассуждении и извлечении [7, 10].

Формирование онтологии – это процесс получения знаний путем преобразования или отображения неструктурированных, полуструктурированных и структурированных данных в экземпляры онтологий. Модуль извлечения информации (*IE*) выполняет большую часть работы путем извлечения структурированной информации из неструктурированных текстовых описаний. Имея выход модуля *IE*, процесс формирования онтологии сводится к созданию индивидуальной онтологии для каждого объекта, полученногона этапе извлечения информации. Если модуль *IE* не может извлечь какой-либо атрибут события, система создает экземпляр с пустыми свойствами. Более того, если событие не обнаружено, то создается экземпляр типа неизвестное событие (*Unknown Event*). Неизвестные события не отбрасываются по нескольким причинам. Формирование онтологии не ограничено событиями, извлеченными из модуля *IE* [8-11].

Семантическое индексирование и извлечение. Для поисковой части был адаптирован подход семантической индексации, основанный на индексах *Lucene6*. Идея заключается в расширении традиционного полнотекстового индекса с извлеченными и выведенными знаниями и изменении рейтинга, чтобы документы, содержащие онтологическую информацию, получали более высокие оценки [9, 11].

Структура семантического индекса имеет первостепенное значение в производительности поиска. Как мы уже упоминали в предыдущих разделах, каждое событие имеет свои связанные с ним свойства, такие как темы и объекты. Эта информация также включается в каждое событие. Мы также включаем полнотекстовые описания, связанные с событиями, в

индекс. Это особенно важно, если тип события неизвестен. Добавление полнотекстовых описаний в индекс допускает неполную информацию о событиях, таким образом, обеспечивает по крайней мере значения отзыва традиционного полнотекстового поиска.

Поиск и ранжирование. При традиционном поиске по ключевым словам индексированные документы обычно содержат только необработанный текст, связанный с этим документом. *Lucene* может легко справиться с такими индексами, и его рейтинг по умолчанию дает обычно хорошие результаты. Тем не менее, сложные индексы должны быть обработаны тщательно. Чтобы воспользоваться преимуществами нашей структуры индекса, используемой в онтологии, мы немного изменили механизм запросов и ранжирования по умолчанию *Lucene*. Прежде всего, мы повысили рейтинг полей, содержащих извлеченную и выведенную информацию, чтобы подчеркнуть их важность. Во-вторых, эти поля повторно оцениваются в соответствии с их важностью [11].

Мы представили новую концепцию семантического поиска, которая включает в себя все аспекты *SemanticWeb,* а именно построение онтологии, извлечение информации, отображение онтологий, логический вывод, семантические правила, семантическое индексирование и поиск.

Вывод поискового робота представляет собой набор документов (для каждого документа задан набор наиболее релевантных понятий), обнаруженные метаданные в соответствии с выбранной онтологией и предложения по развитию онтологии.

На основе пользовательского запроса запускается процесс обхода, в результате которого создается начальный набор извлеченных документов. Полученные документы предварительно обрабатываются с использованием модуля предварительной обработки. Предварительная обработка разделяется на несколько этапов, которые примерно

различаются при извлечении и проверке метаданных RDF в отношении онтологии, обработки и нормализации текста и извлечения гиперссылок.

Предварительно обработанные сегменты документа служат входными данными для процесса вычисления релевантности, который расширяет список URL-адресов для дальнейшей обработки списка документов и контейнера метаданных RDF. Теперь пользователь может проверять результаты процесса обхода, добавлять RDF-метаданные в локальную систему и совершенствовать эволюционирующую онтологию на основе анализа документов, содержащихся в списке документов.

Отображение онтологий– это процесс, в котором семантические отношения определяются между двумя онтологиями на концептуальном уровне, которые, в свою очередь, применяются на уровне данных, трансформируя экземпляры онтологии источника в экземпляры онтологии цели. Здесь отображение онтологии сталкивается с новыми проблемами в контексте *SemanticWeb,* особенно в отношении гетерогенности, динамики, распределения и ограничений в технологии представления.*MAFRA*– это концептуальное описание процесса отображения онтологии, в котором идентифицируются и описываются его фазы [10].

ВЫВОДЫ

По мере развития сети Интернет и появления технологий *SemanticWeb,* становится актуально проблема хранения и поиска релевантной информации. Разрабатываемые поисковые системы в большинстве своем основываются на поиске по ключевым словам, что дает множество неточной информации. Формирование семантических поисковых систем на основе онтологий позволяет решить ряд проблем, связанных с неточностью найденной информации.

В данной статье предлагается архитектура семантической поисковой системы на основе отображения онтологии. В разрабатываемой системе

большое внимание уделено индексированию документов и ранжированию результатов, что дает более точный ответ на запрос пользователя, чем системы, использующие традиционные поисковые механизмы.

Литература

1. Bova V.V., Kravchenko Y.A., Kureichik V.V. Development of Distributed Information Systems: Ontological Approach // Software Engineering in Intelligent Systems. Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015). Vol. 3. – Springer International Publishing AG Switzerland, 2015. – P. 113-122.

2. Bova V.V., Kravchenko Y.A., Kureichik V.V. Decision Support Systems for Knowledge Management // Software Engineering in Intelligent Systems. Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015). Vol. 3. – Springer International Publishing AG Switzerland, 2015. – P.123-130.

3. Dukkardt, A.N., Lezhebokov, A.A., Zaporozhets, D. Informational system to support the design process of complex equipment based on the mechanism of manipulation and management for three-dimensional objects models // Advances in Intelligent Systems and Computing. – 2015. – Vol. 347. – P. 59-66.

4. Кравченко Ю.А. Способы интеллектуального анализа данных в сложных системах / Ю.А. Кравченко, Д.Ю. Запорожец, А.А. Лежебоков // Российская академия наук. Научный журнал. Известия КБНЦ РАН. – Нальчик: Изд-во Кабардино-балкарского научного центра РАН, 2012. – №3 (47). – С.52-57.

5. Кравченко Ю.А., Бова В.В. Нечеткое моделирование разнородных знаний в интеллектуальных обучающих системах // Открытое образование 4(99)/2013. Научно – практический журнал.- M.:CAPITALPRESS, 2013. - С. 70-74.

6. Кулиев Э.В., Новиков А.А., Самойлов А.Н., Старкова Ю.С. Ранжирование онтологий в SemanticWeb / Информатизация и связь, – №3'2016 – С. 97-101.

7. Кравченко Ю.А. Синтез разнородных знаний на основе онтологий // Известия ЮФУ. Технические науки. – 2012. – № 11 (136). – С. 141-145.

8. Старкова Ю.С., Цырульникова Э.С., Кулиева Н.В., Самойлов А.Н. Поисковые системы Semantic Web // Информатизация и связь. 2017. № 3. С. 70-74.

9. Старкова Ю.С., Цырульникова Э.С., Лебединский А.Е. Использование технологии Semantic Web для поиска и ранжирования знаний // Студенческая наука для развития информационного общества. Сборник материалов V Всероссийской научно-технической конференции. . 2016. С. 493-496.

10. Самойлов А.Н., Кулиев Э.В., Новиков А.А., Старкова Ю.С. Поиск и ранжирование знаний в Semantic Web // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2016. № 123. С. 620-629.

11. Новиков А.А., Кулиев Э.В., Самойлов А.Н. Когнитивная архитектура агентов мультиагентной системы // Информатизация и связь. 2016. № 2. С. 127-131.

REFERENCES

1. Bova V.V., Kravchenko Y.A., Kureichik V.V. Development of Distributed Information Systems: Ontological Approach // Software Engineering in Intelligent Systems.

Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015). Vol. 3. – Springer International Publishing AG Switzerland, 2015. – P. 113-122.

2. Bova V.V., Kravchenko Y.A., Kureichik V.V. Decision Support Systems for Knowledge Management // Software Engineering in Intelligent Systems. Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015). Vol. 3. – Springer International Publishing AG Switzerland, 2015. – P.123-130.

3. Dukkardt, A.N., Lezhebokov, A.A., Zaporozhets, D. Informational system to support the design process of complex equipment based on the mechanism of manipulation and management for three-dimensional objects models // Advances in Intelligent Systems and Computing. – 2015. – Vol. 347. – P. 59-66.

4. Kravchenko Ju.A. Sposoby intellektual'nogo analiza dannyh v slozhnyh sistemah / Ju.A. Kravchenko, D.Ju. Zaporozhec, A.A. Lezhebokov // Rossijskaja akademija nauk. Nauchnyj zhurnal. Izvestija KBNC RAN. – Nal'chik: Izd-vo Kabardino-balkarskogo nauchnogo centra RAN, 2012. – №3 (47). – S.52-57.

5. Kravchenko Ju.A., Bova V.V. Nechetkoe modelirovanie raznorodnyh znanij v intellektual'nyh obuchajushhih sistemah // Otkrytoe obrazovanie 4(99)/2013. Nauchno-prakticheski jzhurnal.- M.:CAPITALPRESS, 2013. - S. 70-74.

6. Kuliev E.V., Novikov A.A., Samojlov A.N., StarkovaYu.S. Ranzhirovanie ontologij v Semantic Web / Informatizacija i svjaz', – №3'2016 – S. 97-101.

7. Kravchenko Ju.A. Sintez raznorodnyh znanij na osnove ontologij // Izvestija Sfedu. Tehnicheskie nauki. – 2012. – № 11 (136). – S. 141-145.

8. Starkova Yu.S., Tsyrul'nikova E.S., Kulieva N.V., Samoylov A.N. Poiskovye sistemy Semantic Web // Informatizatsiya i svyaz'. 2017. № 3. S. 70-74.

9. Starkova Yu.S., Tsyrul'nikova E.S., Lebedinskiy A.Ye. Ispol'zovaniye tekhnologiy Semantic Web dlya poiska i ranzhirovaniya znaniy // Studencheskaya nauka dlya razvitiya informatsionnogo obshchestva. Sbornik materialov V Vserossiyskoy nauchno-tekhnicheskoy konferentsii. , 2016. S. 493-496.

10. Samoylov A.N., Kuliyev E.V., Novikov A.A., Starkova Yu.S. Poisk i ranzhirovaniye znaniy v Semantic Web // Politematicheskiy setevoy elektronnyy nauchnyy zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2016. № 123. S. 620-629.

11. Novikov A.A., Kuliev E.V., Samoylov A.N. Kognitivnaya arkhitektura agentov mul'tiagentnoy sistemy // Informatizatsiya i svyaz'. 2016. № 2. S. 127-131.