УДК 004.822

В.И. Воробьев, Е.Л. Евневич

Поиск сервисов в облачных вычислениях

Аннотация. Рассмотрены методы семантического поиска сервисов на основе онтологий и тезаурусов. Проведен обзор отечественных и зарубежных методик. Предложена система поиска на основе вероятностных онтологий.

Ключевые слова и фразы: таксономия, онтология, метаданные, контекст, словоформа.

Введение

Рост количества IT-решений в виде сервисов на основе облачной архитектуры часто приводит к ситуации, когда в крупных компаниях количество внутренних сервисов может исчисляться сотнями, а в случаях крупных IT-компаний даже тысячами. При этом дополнительная специфика ситуации состоит в том, что многие подразделения одной и той же организации часто занимаются разработкой и внутренним использованием собственных сервисов, образуя при этом собственную облачную сервисную систему. В такой ситуации острым становится вопрос поиска нужных сервисов. Подобный поиск можно реализовать при помощи онтологической таксономии, в которую включены нужные компоненты. При этом обратный индекс поисковой машины строится на основе метаописания сервиса, а таксономия облегчает пользователю выбор конечного решения. Такая онтологическая таксономия может предоставить визуальное древовидное представление структуры, в которой для пользователя выделены требуемые узлы: сервисы, подходящие под его запрос. Таким образом, сочетание стандартного поискового индекса и онтологической таксономии могут значительно ускорить и упростить поиск требуемого сервиса в сложной внутрикорпоративной IT-структуре.

Как правило, каждый облачный модуль сбора и обработки данных сопровождается подробной текстовой документацией и метаописанием. Поэтому поиск сервиса может рассматриваться как поиск текстового документа с несколькими полями: документация интерфейса программирования (API) и мета-документация сервиса, в которой описана функциональность, типы входных и выходных данных и специфика их обработки.

1. Обзор поисковых методов

Качество стандартного поиска текстовых документов с помощью одного поискового индекса принято улучшать с помощью стандартных методов. Один из таких методов – расширение и переформулировка пользовательского запроса с помощью тезауруса. Тезаурус представляет собой специальный словарь, в котором содержатся кольца семантически близких концепций (термов). Если в запросе пользователя встретился один из термов кольца, то запрос расширяется всеми остальными термами из этого кольца. Например, запрос «обработка статистики дорожного движения» будет расширен термами «траффик, автотранспорт, транспортный поток» и т.д. Таким образом, за счет повышения полноты поиска улучается его качество.

Существует множество подходов к извлечению семантических отношений и автоматической генерации тезаурусов. Все эти подходы так или иначе исследуют внешние источники знаний на предмет наличия отношений. Одна группа методов использует распределение терминов в коллекции документов. Например, авторы работы [1] используют максимальное расстояние между документами, содержащими термы с целью группирования термов в кластеры. Расстояние в данном

случае определяется как расстояние между двумя векторами в модели bag-of-words (документ представляется в виде набора слов без учета их положения и связей в тексте).

Другая группа методов использует гипотезу о распределении и строит кластеры связанных термов, основываясь на близости контекстов, в которых они встречаются. Примером применения такого подхода может служить работа [2]. Особенностью этой работы является различие контекстов по грамматике, например, авторы отделяют контексты, в которых терм является субъектом, от тех, в которых тот же терм появляется в качестве объекта.

Третий подход использует знания о структуре ссылок в гипертекстовой разметке. Два наиболее ярких проекта, служащих источником информации о семантических отношениях - это WordNet и Википедия. Пример использования WordNet описан в работе [3]. Статья [4] описывает использование Википедии для поиска связанных сущностей. Основная разница между ними состоит в том, какие типы ссылок анализируются, например, ссылки по категориям [5] или ссылки внутри статьи [6].

Более современные подходы основываются на декомпозиции матриц, например, латентно-семантический анализ, описанный в статье [7], может применяться в связке с вышеописанными методами.

Также следует отметить другое направление по генерации тезаурусов, использующее явную синонимию и другую семантическую информацию, содержащуюся в тексте. Первой в этой области была работа [8]. Этот подход успешно применялся во многих проектах, включая извлечение синонимов биомедицинской тематики [9], построение классификатора отношений "часть-целое" [10] и т.д. Этот метод используется в данной работе в сочетании с заранее определенным набором лексических паттернов для извлечения синонимических названий продуктов.

С учетом неэффективности синтаксического поиска во многих случаях, в данной работе предлагается способ семантического поиска на основе использования базы данных, а именно множества семантических классов русского языка. Персонализированный семантический поиск, навигация и методы получения и анализа контента представляют интерес как средство сужения поиска и повышения релевантности результата. Для осуществления смыслового поиска в web-документах требуется предварительный семантический анализ их содержимого с целью улучшения структурирования информации в документах. При переходе к семантическому описанию документов происходит сжатие и обобщение информации, что приводит к новому знанию. В то же время семантическое представление информации невозможно без определения закономерностей совместного употребления слов, где значения каждого слова определяется контекстом его использования. Проблема вычислимости здесь связана с тем, что значение каждого из слов этого множества в свою очередь определяется собственным контекстом, состоящим из слов, в число которых может попасть и исходное.

Еще один подход к семантическому поиску подразумевает построение некоторой математической модели языка. Тогда любое слово в русском языке можно рассматривать как имя функции, где последняя в качестве результата возвращает семантику слова [11]. При этом конкретное значение слово получит только после подстановки аргументов (это не обязательно контекст слова), а его смысл будет вычислен по мере выполнения функции. Предложение в данном случае — это законченная суперпозиция функций, а смысл предложения вычисляется при построении и выполнении этой суперпозиции. Такой подход к семантическому анализу позволяет в том числе построить онтологии, описывающие предложения. Кроме того, онтологии, как и тезаурус, можно использовать еще на этапе семантического анализа. В этом случае наиболее эффективно их совместное использование, где онтология описывает комплекс понятий и отношений предметной области, а тезаурус формирует подобную систему понятий и отношений, но в рамках лингвистических знаний по предметной области [12].

2. Поиск на основе словоформ русского языка.

Основная задача любой поисковой системы состоит в приведении множества поисковых запросов к множеству поисковых ответов. Поисковыми запросами для нее являются пользовательские или клиентские запросы (при этом в некоторых случаях клиентами могут быть другие поисковые системы), а поисковыми ответами - информационные представления web-документов (обычно это URL-адрес и фрагмент текста), в которых производился поиск. Традиционные системы синтаксического поиска (например, Google) решают упомянутую задачу при помощи различных статистических алгоритмов. Система семантического поиска также использует статистический алгоритм, однако в основе ее работы лежит не он, а база, содержащая в себе некоторое представление знаний. К такой базе можно отнести любое множество данных, если оно содержит в себе ту или иную лингвистическую информацию (слова, термины...) и подготовлено (формализовано) для адекватной и однозначной интерпретации вычислительной машиной. Это могут быть электронные словари, классификации, базы данных и др. Это не может быть текст на естественном языке, так как сам по себе он не может быть однозначно интерпретирован машиной. Стоит отметить, что такой текст также не всегда может быть однозначно интерпретирован и человеком из-за многозначности слов. В данной реализации в качестве базы с представлениями знаний использован словарь Тузова В.А. [13,14], содержащий в себе информацию о соответствии между множеством слов (точнее словоформ для каждого из слов) на русском языке и множеством семантических классов. Семантический класс представляет собой некоторую сущность или понятие, имеющие определенный смысл для человека. Возьмем, к примеру, класс "Огонь". Ему могут соответствовать слова "гореть", "поджигать", "обжечь", "тушить" и др. Одному слову может быть поставлено в соответствие несколько семантических классов. Например, слово "кинофестиваль" может быть связано с классами "Кино" и "Фестиваль".

Система семантического поиска подразделяется на две основные подсистемы: подсистему сбора и анализа информации, и подсистему поиска. Подсистема сбора и анализа информации представляет собой программный канал, принцип работы которого соответствует принципу работы «pipe-line» в системе Unix.

Представление неопределенности, присущей объектам и явлениям реального мира, требует применения не традиционных, а вероятностных онтологий. При создании профиля применяется вероятностная онтология идентификации пользователя. В качестве шаблона используется референтная архитектура для разработки вероятностных онтологий. Референтная архитектура для разработки вероятностных онтологий (RAPOD) каталогизирует и определяет процессы и объекты, необходимые для разработки, внедрения и оценки вероятностных онтологий, разработанных для обмена знаниями и неоднократного использования в предметной области. Онтологии также обеспечивают системную интероперабельность. Они создают представление семантики области, поддающееся машинной интерпретации, предоставляя таким образом возможность обмена информацией с четко определенным значением.

Заключение

Характерным свойством предложенной системы, интересным для дальнейшего исследования, является неограниченная длина поискового запроса, что позволяет осуществлять поиск и сравнение текстов по текстам. Теоретически увеличение размера текста увеличивает точность его семантического анализа и результатов поиска. Для практического использования системы необходимо провести ее объемное тестирование на основе формальных критериев оценки

результатов семантического поиска. Кроме того, возможно дополнение существующей базы данных и поиск новых на основе существующих и собственных статистических алгоритмов и метрик.

Список литературы:

[1] Carolyn J. Crouch and Bokyung Yang. *Experiments in automatic statistical thesaurus construction.* In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '92, pages 77–88, New York, NY, USA, 1992. ACM.

[2] Takenobu Tokunaga, Iwayama Makoto, and Tanaka Hozumi. *Automatic thesaurus construction based on grammatical relations,* 1995.

[3] Edward A. Fox, J. Terry Nutter, Thomas Ahlswede, Martha Evens, and Judith Markowitz. *Building a large thesaurus for information retrieval.* In Proceedings of the second conference on Applied natural language processing, ANLC '88, pages 101–108, Stroudsburg, PA, USA, 1988. Association for Computational Linguistics.

[4] David Milne, Olena Medelyan, and Ian H. Witten. *Mining domain-specific thesauri from wikipedia: A case study*. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, WI '06, pages 442–448, Washington, DC, USA, 2006. IEEE Computer Society.

[5] Michael Strube and Simone Paolo Ponzetto. *Wikirelate! computing semantic relatedness using wikipedia.* In proceedings of the 21st national conference on Artificial intelligence - Volume 2, pages 1419–1424. AAAI Press, 2006.

[6] Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. *Wikipedia mining for an association web thesaurus construction.* In Proceedings of the 8th international conference on Web information systems engineering, WISE'07, pages 322–334, Berlin, Heidelberg, 2007. Springer-Verlag.

[7] Tonio Wandmacher. *How semantic is latent semantic analysis?* In Proceedings of TALN/RECITAL.

[8] Marti Hearst. *Automatic acquisition of hyponyms from large text corpora.* In Proceedings of the Fourteenth International Conference on Computational Linguistics.

[9] John McCrae and Nigel Collier. *Synonym set extraction from the biomedical literature by lexical pattern discovery.* BMC Bioinformatics, 9(1):159, 2008.

[10] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. *Learning syntactic patterns for automatic hypernym discovery.* In NIPS, 2004.

[11] J. Gonzalo, F. Verdejo, I. Chugur, and Cigarran J. *Indexing with wordnet synsets can improve text retrieval.* In Proceedings of the COLING/ACL '98.

[12] Нариньяни А. С. *Кентавр по имени ТЕОН: Тезаурус + Онтология.* Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Аксаково, 2001. Т. 1. С. 184–188.

[13] Perminov, S.V.; Vorobyev,V.I.; Atiskov, A.J. Declarative *transformation of arbitrary structured data to ontologies.* EUROCON 2009, EUROCON '09. IEEE 18-23 May 2009. - P. 426-431.

[14] Тузов В. А. *Компьютерная семантика русского языка* СПб.: Издательство Санкт-Петербургского университета, 2004.

[15] Воробьев В.И., Перминов С. В., Атисков А.Ю. *Декларативные способы преобразования исходных данных в онтологии.* Труды симпозиума «Онтологическое моделирование: состояние и направления исследований и применения», М.: ИПИ РАН, 19 – 20 мая, 2008 г., с. 267-277.