УДК 519

Современные проблемы развития технологии семантического поиска

А.Ю. Дорофеева
АлтГУ, г. Барнаул

Семантическая паутина (Semantic Web) является расширением традиционного Интернета и нацелена на упрощение поиска и распределения информации. Технология семантического поиска основывается на элементах, построенных с использованием стандартных языков онтологий, таких как OWL (Ontology Web Language). Обычные поисковые системы основываются на поиске ключевых терминов запроса в документе и не могут использовать его смысловое значение для получения результата, поэтому сообщество исследователей семантической паутины предложило использовать семантические поисковые технологии, среди которых OntoSearch, Semantic Portals, Semantic Wikis, мультиагент P2P, семантические системы маршрутов (запросов), вопросно-ответные системы, использующие онтологии для хранения баз знаний [1].

Документ семантической паутины SWD (Semantic Web Document) можно рассматривать как набор данных, контентом которого является либо онтология, либо обычный документ, размеченный определенными тегами, взятыми из онтологии предметной области. Такие Интернет-документы могут быть распределены по множеству различных категорий, относящихся к типам онтологий, используемых для разметки документа. Примерами таких категорий являются тяжеловесная или легковесная онтологии.

Хотя семантическая паутина способствует поиску информации в сети, существует несколько нерешенных проблем, которые следует принять во внимание. Первая из них – это огромное количество неструктурированных Интернет-документов, которые должны быть семантически размечены для использования семантическими поисковыми системами. Это непростая задача, т. к. она, среди прочего, требует развития проблемно-ориентированных онтологий.

Полностью автоматизированный процесс разметки существующих данных – еще одна нерешенная задача. С другой стороны, эффективный поиск Интернет-документов требует, вне существования онтологий, создания формальных запросов. Получается, что обычные пользователи Интернета должны изучить формальный язык для создания та-

кого рода запросов, а это не так просто. Методы, позволяющие автоматизировать процесс преобразования запросов свободной формы (например, в форме предложения на естественном языке или как множество/список ключевых слов) к формальному виду, в настоящее время являются объектом исследования. Построение отображения онтологий предметных областей на формальные запросы также активно исследуется.

Кроме того, при разработке и реализации семантических поисковых систем возникает еще ряд проблем:

1. Использование внешних ресурсов.
2. Автоматизация и прозрачность.
3. Производительность.
4. Точность/полнота.

Исследования мировых лидеров в прогнозных исследованиях (IDC, 2012; Tofler, 2006; GARTNER, 2012) показывают, что до 2020 года количество информации и потребности в ней будут расти экспоненциально. Таким образом, одной из самых больших проблем современного общества является информационное переполнение. Как представляется, базовые тренды в области семантических технологий в значительной мере связаны с концепцией Semantic Web, которую в 2000 г. выдвинул Тим Бернерс-Ли – один из основоположников WWW.

С момента появления этой концепции прошло уже более 10 лет, но пока SW-эра, в отличие от эпохи Интернет, еще только приближается и на этом пути существует значительное число научных, технических, технологических и человеческих проблем, основными из которых являются доступность семантического контента, доступность онтологий и средств их разработки, а также эволюция онтологий, масштабируемость, мультиязыковость, визуализация и стабильность.

Доступность семантического контента является основной проблемой на пути формирования и использования пространств знаний, так как сейчас основная масса информации не представлена в «семантических» форматах и нет надежды, что эта работа может быть выполнена вручную. Онтологии, по мнению практически всех специалистов, являются ключевым компонентом в решении проблемы семантизации контента. В связи с этим особое значение приобретают проблемы онтологического инжиниринга, а также доступность уже существующих онтологий.

В настоящее время основными используемыми системами информационного поиска являются представители традиционных подходов поиска по ключевым словам (Google, Bing, Yahoo и т.д). Примерами применения семантических технологий в глобальных масштабах мож-

но считать проекты SearchMonkey от Yahoo, Rich Snippets от Google, или Bing Powerset. Такие системы подтверждают тот факт, что семантический поиск является перспективным направлением развития поиска информации, и решение проблем семантического поиска откроет новые возможности как для отдельно взятых компаний, так и человечества в целом.

Библиографический список

1. Allemang D., Hendler J. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL // Morgan Kaufmann, 2008.

Bernstein A., Kaufmann E., Fuchs N. Talking to the semantic web – a controlled english query interface for ontologies // AIS SIGSEMIS Bulletin. – 2005. – № 2. – P. 42–47.

2. Corcho O. Ontology based document annotation: trends and open research problems // Int. J. Metadata, Semantics and Ontologies. – 2006. – № 1. – P. 47–57.

3. Хорошевский В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1) / В. Ф. Хорошевский // Искусственный интеллект и принятие решений. – 2008. – № 1.

УДК: 519.237.8

Описание одного алгоритма кластеризации типа Forel

В.В. Журавлева, А.А. Бондарева
АлтГУ, г. Барнаул

В научной литературе описано множество методов автоматической классификации. Среди них есть достаточно простые (k-средних, Forel), применение которых позволяет построить классы простой формы, и трудоемкие (KRAB), но строящие при этом «необычные» классы [1, 2]. Здесь описан простой алгоритм, позволяющий получать классы необычной формы.

Предварительно стоится разбиение массива данных (представленного таблицей числовых значений признаков, в которой строки соответствуют объектам, а столбцы признакам) на большое количество мелких непересекающихся кластеров с помощью алгоритма Форел. Также для начального разбиения можно использовать иной алгоритм. В итоге каждый кластер можно описать сферой определенного радиуса. Суть алгоритма состоит в последовательном объединении маленьких сфер-кластеров в большие классы по принципу «ближайшего соседа».