УДК 004.8

Н.Н. Гдибовец, Л.И. Глибовец, А.С. Шабинский

ПРИМЕНЕНИЕ ОНТОЛОГИИ И МЕТОДОВ ТЕКСТОВОГО АНАЛИЗА ПРИ СОЗДАНИИ ИНТЕЛЛЕКТУАЛЬНЫХ ПОИСКОВЫХ СИСТЕМ

Введение

С момента становления информационного поиска (ИП) как отдельного научного направления его основной проблематикой неизменно является качественное удовлетворение информационной потребности пользователя. С появлением концепции семантического веба информационный поиск развивается в направлении систем с имитацией понимания сути информации [1]. Пользователь заинтересован не просто найти множество документов, отвечающих его запросу (в идеале информационной потребности), а и отличить документы, соответствующие запросу. Для этого в поисковых системах вычисляются точные показатели, а результирующее множество документов упорядочивается в соответствии с выбранной методикой ранжирования.

Пользователей и исследователей интересует способность поисковой системы находить релевантные документы. Эффективность системы ИП оценивают бинарной классификацией документов коллекции на релевантные и нерелевантные, а релевантность — как информационную потребность пользователя, а не простое сооткетствие между запросом и найденным локументом. Важная задача ИП — обеспечение достаточной выразительности языка запросов, не менее важны задачи классификации и кластеризации документов.

В настоящее время в качестве формы представления знаний чаице применяются онтологии, Онтология 8 философии — наука о бытии, в информатике — явная спецификация концептуализации, лежашей *в основе формального* представления знаний как абстрактного представления реальности, которую необходимо отразить для некоторых потребностей [2]. Так, в [3] онтология определена как спецификация концептов *естественного* ***или*** формального языка; отношение между концептами; правила применения отношений к концептам; интерпретация каждой спецификации концепта в соответствующее понятие предметной области.

По ряду требований интеллектуальные поисковые системы (ИПС) преобладают над классическими поисковыми системами, основанными на поиске по ключевым словам. Преимущества проявляются не только в модели ИП, но и во взаимодействии с пользователем, в участии экспертов, в механизме ранжирования результатов. В ИПС сделана попытка «понять» информационную потребность пользователя, учитывая, что пользователь часто не может четко сформулировать информационный запрос к ИПС. В ИПС накопление знаний на основе опыта, в частности экспертов, трансформирует поиск в семантический по источникам из Интернет.

Образцом ИПС является система, в которой определение соответствия документа информационной потребности (а вместе с этим и ранжирование) выполняет человек, а остальные необходимые процессы ИП (сбор, индексация, классификация, любая механическая обработка информации и т.п.) возлагаются на программный вычислитель. Такая ИПС гипотетическая, но она лучше передает суть идеализированной поисковой системы, в которой информационная потребность пользователя удовлетворяется в полной мере, естественно, при наличии нужной информации.

1. Использование онтологии для определения идентичности документов в модели eTVSM

Улучшенная тематическая векторно-пространственная модель ИП (Enhanced Topic-based Vector Space Mode! — eTVSM) [4] эволюционировала из тематической векторно-пространственной модели (Topic-based Vector Space Model — TVSM), которая, в свою очередь, эволюционировала из классической векторно-пространственной модели (Vector Space Mode) — VSM).

1.1. Векторно-пространственная модель. Впервые она использовалась в 1960-х годах в поисковой системе SMART (System for the MechanicaJ Analysis and Retrieval of Text) Information Retrieval System, разработал ее Дж. Салтон [4]. Она относится к классу алгебраических и представляет документ в виде вектора в многомерном векторном пространстве. Каждое измерение пространства соответствует одному термину, поэтому главный постулат модели — ортогональность (независимость) терминов. Компоненты вектора соответствуют терминам в документе, а значение компонента определяется значимостью (весом) термина в документе. Вес термина рассчитывается на основе таких статистических метрик, как частота появления в тексте и т.д. Наконец, соответствие между документами рассчитывается с помощью скалярного произведения и часто выражается косинусом угла между векторами.

Векторно-пространственную модель неоднократно критиковали за ее специфичность, неприспособленность к большим объемам информации из-за малых скалярных произведений и большой размерности векторов; невосприимчивость к различным словоформам; за наличие подстрок, способных привести к ошибочной релевантности (ошибочная релевантность в информационном поиске означает наличие в результирующей выборке таких документов, которые на самом деле не являются релевантными к запросу (в статистике — ошибка второго рода)); семантическую невосприимчивость, за различные термины с одинаковым или похожим неассоциируемым значением, вызывающим ошибочную нерелевантность (ошибочная нерелевантность означает, что релевантные документы не попадают в результат (в статистике — *ошибка первого рода)) и* игнорирование порядка терминов в документе.

1.2. Тематическая векторно-пространственная модель, В модели TVSM преодолены ограничения VSM [5] благодаря введению отношений между словами. Поначалу улучшения модели достигли, устранив предположение об ортогональности терминов, затем ввели понятие фундаментальных тем как векторов в ортогональном базисе векторного пространства. Так, принцип независимости «перешел» от терминов *к* темам. Вес термина определен как модуль вектора:

$$\vec{t}_k = (t_{k1}, t_{k2}, \ldots, t_{kd}) \in R,$$

t_k — термин, R — векторное пространство, d — количество тем (размерность пространства R):

$$|\vec{t}_k| = \sqrt{t_{k1}^2 + t_{k2}^2 + \ldots + t_{kd}^2} \in [0; 1].$$

При таком представлении тем и терминов направление вектора-термина определяет принадлежность термина теме или темам, с которыми он имеет одинаковое направление, причем мера принадлежности определяется углами к соответствующим векторам-темам. Для тематически неспецифических терминов (союзы и т.д.) угол равнозначен всем векторам-темам и равен 45°. Аналогично длине векторов-терминов тематически специфические термины имеют длину, стремящуюся к единице, неспецифические — к нулю.

ISSN 0572-2691

В модели TУ8М документы тоже представлены векторами. Векторы-документы *для* удобства нормализованы к единице по длине:

$$\vec{d}_k = \frac{\vec{\delta}_k}{|\vec{\delta}_k|} \in R, \quad \vec{\delta}_k = \sum_{i=1}^{n} e_{ki}\vec{t}_k,$$

где e_{ki} — частота появления термина i в документе k.

Сходство двух документов определяется по скалярному произведению, предложенному в У5М. Благодаря нормализации векторов-документов сходство документов является косинусом угла между соответствующими векторами:

$$\mathrm{sim}(d_1, d_2) = \vec{d}_1\vec{d}_2 = |\vec{d}_1||\vec{d}_2|\cos\alpha = \cos\alpha.$$

Поскольку модель TVSM отражает лишь общую идею представления документов и определения сходства между ними, в ней не определены фиксированные методы задания углов и веса векторов, а лишь некоторые требования для повышения эффективности.

Вес термина, специфический для определенных тем, близок к единице; вес неспецифических — к нулю. Угол между словами с общей основой должен быть 0°. Угол между синонимами или *словами,* близкими по теме, должен приближаться к 0°, а между тематическими словами и стоп-словами (союзы, предлоги, артикли) быть 45 Угол между стоп-словами должен быть близок к 0

По сравнению с классической моделью VSM модель TVSM повышает восприимчивость системы к связям между словами, в частности синонимии, и словоформам (падежам, множественному числу, склонениям и т.п.)-

Модель TVSM эволюционировала в улучшенную TVSM, в которой детализирована концепция отношений между понятиями благодаря устранению независимости между темами и использованию онтологии как источника знаний о семантической связи между понятиями предметных областей. В eTVSM способ определения сходства документов построен не на принципе сходства терминов, а на основе их интерпретации. Модель оперирует понятиями слова, основы слова, термина, интерпретации и темы. Документы представлены в виде векторов-интерпретаций [6, 7J.

1.3, Улучшенная тематическая векторно-пространственная **модель,** В eTVSM онтология использована для формирования операционного векторного пространства на основе понятий термина, его интерпретации и темь/ ***[6].*** Устранен принцип независимости тем и ортогональность соответствующих векторов, зато отношения между темами, заложенные в онтологии, прежде всего определяют углы между соответствующими векторами, используя карту тем (topic map). Карта тем — ориентированный граф, узлами которого являются темы, а ребрами — отношения «супертема-субтема». Наконец, сходство тем расчетывается как скаляр-Ное произведение векторов-тем [7].

В модели eTVSM интерпретации используются как промежуточные звенья между темами и терминами и несут семантическую нагрузку. Интерпретации дают свободу разработчикам поисковых систем определять связи между понятиями в зависимости от целей или потребностей. Формализм расчета сходства интерпретаций аналогичен темам [6]:

$$\vec{\phi}_i = \frac{w(\phi_i)}{\left|\sum_{\tau_k \in T(\phi_i)} \vec{\tau}_k\right|} \sum_{\tau_k \in T(\phi_i)} \vec{\tau}_k, \qquad (1)$$

где $\phi_i \in \Phi$ — множество всех интерпретаций, $w(\phi_i) \in [0;1]$ — вес интерпретаций, $T(\phi_j) \in \vartheta(\Theta)$ — множество тем соответствующей интерпретации, $\vartheta(\Theta) = 2^{\Theta}$, Θ — число тем.

Термин как единица информации в eTVSM может состоять из нескольких слов и обозначать целостное понятие. С каждым термином связано произвольное количество интерпретаций. Между самими терминами невозможны связи, что устраняет циклы из структуры онтологии [6, 7].

В отличие от предыдущих систем, eTVSM оперирует интерпретациями как основными носителями содержания документов вместо работы с исходными текстами документов. Восприимчивость практически ко всем лингвистическим связям и семантическим отношениям между понятиями существенно улучшает ИП.

Путь от документа к интерпретации преодолевается в несколько этапов [6].

Документ — Простой текст. Из исходного текста устраняется форматирование и метаданные.

Простой текст —> Снова. Путем токенизации (tokenization) пробелами текст разбивается на отдельные слова.

Слова —> Основы слов. Стемминг-алгоритм (стеммер) преобразовывает слова из входных словоформ в начальную форму, отсекая суффиксы и окончания и оставляя основу слова. Результат зависит от конкретной реализации стеммера.

Слова -> Понятия. На основе онтологии из набора слов выбираются отдельные понятия, присущие онтологии. Устраняются стоп-слова, не принадлежащие определенным терминам.

Понятия — Интерпретации. На основе онтологии сопоставляются понятия интерпретации.

Подобие документов в eTVSM определяется аналогично TVSM, с той лишь разницей, что вместо векторов-тем и веса тем используют векторы-интерпретации и вес интерпретаций [6]; d_i e O, где D— множественное число всех документов,

$$\forall d_i \in D : \vec{d}_l = \frac{\vec{\delta}_l}{|\vec{\delta}_l|}, \quad \vec{\delta}_l = \sum_{\phi_j \in \Phi} \omega_{ij} \vec{\phi}_j, \text{ где } \omega_{ij} \text{ — вес интерпретации } j \text{ в документе } i.$$

Теоретически модель eTVSM привлекательна для построения высокоэффективных поисковых систем, в основном за счет онтологий, т.е. восприимчивости к семантической связи между понятиями в документах. Очевидно, это существенное преимущество по сравнению с классическими поисковыми системами, где весь процесс поиска основан на ключевых словах и их словоформах. От качества моделирования онтологий зависит эффективность поисковых систем,

В [7] предложен подход к автоматическому построению онтологии в eTVSM на основе WordNet (WordNet — большая лексическая база английского языка, разработанная под руководством Дж. Миллера в Принстонском университете, это сеть содержательно связанных слов и понятий, представленных множеством синонимических существительных, прилагательных, наречий и глаголов). В результате сравнений сделан вывод, что eTVSM с онтологией на основе WordNet менее эффективна, чем eTVSM с онтологией синонимов и даже VSM. Как онтология общего назначения WordNet не может отразить большинство устоявшихся сложных понятий, специфику их значения и контекста. Зато в [6] предложено использовать полуавтоматический подход к моделированию онтологий для eTVSM.

В работе [8] вместо интерпретаций eTVSM использованы аннотации документов, причем понятия в документе сопоставляются с определенными предметными областями. Хотя механизмы реализации интерпретаций* и специфические

* Интерпретация в общем случае, как указано ранее.

ISSN 0572-2691

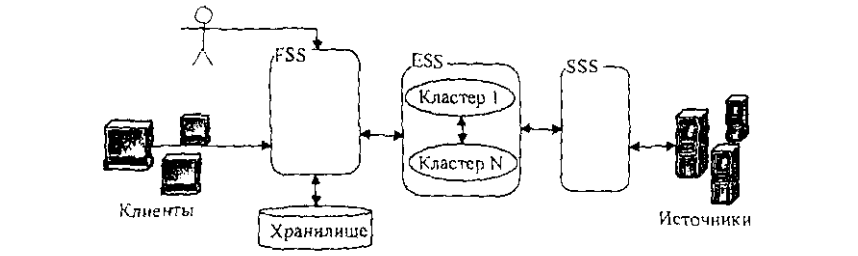
моменты в [7] и [8] существенно отливаются, обший подход остается неизменным и заключается в использовании онтологий в виде специализированных баз знаний для имитации определения смысла понятий и их значений в документе.

2. Использование онтологий для интеллектуального сбора информации

В системах ИП определению соответствия между документами предшествуют эффективный сбор и первичная обработка разрозненной информации во множестве первоисточников. Маловероятно, что сфера заинтересованности пользователя ограничивается лишь отдельным ресурсом, а соединение нескольких источников в один не может быть оптимальным, прежде всего, ввиду права собственности и неоправданного дублирования больших объемов информации, тем более, что в общем случае областью поиска может стать выбор из сотен или даже тысяч электронных ресурсов. Возможности классических поисковых систем (Google, Bing, Yandex) далеки от семантического сбора информации и в основном базируются на непосредственном анализе ключевых слов. Итак, в семантической сети, а в итоге и в ИПС есть задача эффективного интеллектуального сбора информации из большого числа источников.

Рассмотрим онтологически-ориентированный сбор информации с ограниченными предметными областями на примере системы AGATHE [9]. Как совместный проект Франции и Бразилии AGATHE реализует кооперативный подход (Coopérative Information Gathering) к сбору информации на основе программных агентов и онтологий выбранных предметных областей.

Концептуально система состоит из трех подсистем (см. рисунок): внешнего интерфейса (Frontend subsystem — FSS), извлечения информации (Extraction subsystem — ESS) и поисковой подсистемы (Search subsystem — SSS).



Поисковая подсистема отвечает за запросы к внешним информационным источникам и непосредственный сбор информации. Она использует три вида агентов; поисковые, работающие с известными поисковыми системами (Google, Bing); ресурсные, работающие с ресурсными сайтами, и цитатники, например CiteSeer; надзорный агент, контролирующий функционирование агентов двух первых видов.

Полученные поисковой подсистемой документы обрабатывает подсистема извлечения информации как центральный компонент архитектуры AGATHE. Она состоит из нескольких кластеров (extraction cluster), каждый отвечает за определенную предметную область. Агенты этой подсистемы выполняют функции классификации, получая информацию из онтологий предметных областей и единой служебной онтологии AGATHE. Каждый кластер имеет нескольких агентов с разными задачами. Информация, извлеченная из этой подсистемы, передается в подсистемы внешнего интерфейса, управляющие хранением результатов классификации и получением релевантных данных.

Подсистема внешнего интерфейса поддерживает взаимодействие с системой. Кроме непосредственно компонента, отвечающего за контакт с пользователем или с некоторым клиентским приложением, есть компонент, управляющий хранением информации, полученной от подсистемы извлечения информации.

В AGATHE применяется несколько тематических и одна внутренняя онтологии, Тематические онтологии используют агенты соответствующих предметных областей для анализа содержания документов (веб-страниц). Внутренняя служит основой классификации материалов, определяя главные сущности относительно веб-страниц (адрес, резюме, содержание), отдельные понятия в тексте и т.п.

3. Онтологически-ориентированная интерпретация *ключевых слов* для семантического поиска

Обычно пользователь упрощенно, интуитивно выражает свою потребность, часто в виде нескольких ключевых слов. В ИПС необходимо интерпретировать ключевые слова в более выразительные и контекстно-ориентированные запросы в терминах самой системы, а не в терминах пользователя. Оценим онтологически-ориентированный подход к интерпретации ключевых слов [10],

3.1. Модели интерпретации запросов. Традиционный поиск по ключевым словам основан на модели запросов и модели ресурсов, а семантический, базирующийся на онтологиях, — на четырех других моделях [10].

Модель мышления (ментальная модель) O_U формализует информационную потребность пользователя в начале процедуры информационного поиска. Поскольку истинные механизмы человеческого мышления далеки от понимания, то постулируется лишь тот факт, что модель состоит из сущностей, соответствующих предметам реального мира, и принадлежит области знаний пользователя. Естественно, пользователь может интересоваться сущностями, не относящимися к модели и называемыми пробелами (gaps).

Модель вопросов пользователя Q_U состоит из элементов, которые конструируются из языковых примитивов P_U языка пользователя L_U. Эта модель выражает элементы O_U в виде элементов P_U языка пользователя L_U.

Модель системных ресурсов O_S состоит из языковых примитивов P_S языка системы L_S, основана на знаниях, выраженных в онтологиях, и в отличие от абстрактной модели O_U имеет четкую и доступную структуру. Элементы модели формируют базу знаний ИПС.

Модель вопросов системы Q_S представляет окончательно обработанную версию вопроса пользователя как конструкцию из языковых примитивов P'_S языка запросов системы L'_S, которые обычно выражены элементами онтологий ИПС.

3.2. Онтологически-ориентированная интерпретация запроса. В работах [7, 9, 10] четко просматривается наличие единого принципа отражения реальности в интеллектуальных системах интерпретаций. Правда, формально механизмы во всех трех исследованиях различаются, но сохраняют общую природу — сущности реального мира формально соотносятся с сущностями модели знаний. В [10] соотносимость обеспечивается одноименным предположением. В целом подход основан на двух предположениях:

П1 (онтологически-мысленная соотносимость): все сущности O_U семантически и структурно соотносятся с сущностями O_S;

П2 (локальность информационной потребности): все элементы $O'_S \subseteq O_S$, отвечающие конкретной информационной потребности O_U, должны быть связаны цепочкой максимальной длины d:

$$\forall a, b \in O'_S : \langle a, b \rangle \oplus \exists (x_n)_{n=0}^{d}, a = x_0, \langle x_0, x_1 \rangle .. \langle x_{n-1}, x_n \rangle \langle x_n, b \rangle .$$

Предположение предусматривает, что пользователь мыслит структурами, полностью соотносимыми с сущностями онтологий, причем каждая отдельная информационная потребность отображается только в части онтологии.

В терминах декларативной логики П1 переформулирум более строго.

П1′ (допускается, что мыслительные модели пользователей построены подобно базам знаний, основанным на декларативных логиках). Сущности O_U, принадлежащие дизъюнктивному объединению $\coprod\{I, T, C, D, R, U\}$, отвечают сущностям онтологии на основе SHOIN(D) (I — индивиды, T — значения величин, C — понятия (concepts), D — интервалы величин, R — объектные свойства, U — количественные свойства).

Ассоциации в O_U пользователей соответствуют отношениям

$$\langle i, C\rangle, \langle i_1, R, i_2\rangle, \langle i, U, j\rangle,$$

где $i, i_1, i_2 \in I$, $j \in T$.

Приведем схему алгоритма интерпретации ключевых слов [10].

Шаг 1. Термины языка пользователя сопоставляются сущностям в онтологии $f : Q_U \to O_S$.

Шаг 2. Поиск связей между сущностями в онтологии.

Шаг 3. Построение запросов.

Шаг 4. Вычисление всех возможных цепочек между элементами O'_S.

Шаг 5. Сопоставление запросов вычисленным цепочкам.

Шаг 6. Ранжирование запросов (в соответствии с П2 лучшей будет более короткая цепочка).

Итак, в ИПС сначала пользователь формулирует запрос в виде ключевых слов, обработка *которых* возвращает соответствующие онтологические сущности (шаг 1) алгоритма интерпретации ключевых слов. Применяя к ним шаг 2, строится граф всех связанных сущностей с ограниченой длиной цепочки связей. На этом этапе пользователь может выбрать дальнейшее расширение графа (уточнение, детализацию запроса) либо перейти к результатам поиска по запросу. В первом случае он может улучшать запрос в рамках определенной онтологии, во втором запрос передается на следующий этап (шаг 3). По одному из окончательных формальных запросов будет построен результат поиска с использованием базы знаний.

Заключение

В примерах использования онтологий на этапах сбора и классификации информации, интерпретации пользовательских запросов, поиска релевантной информации в документе особый акцент сделан на качественном подборе функции интерпретации. Разработчики систем ИП могут свободно определять *интерпретации* и это оказывает существенное влияние на общую эффективность системы. Рассмотренные нами разные роли онтологий в поисковых системах при создании интеллектуальных поисковых систем, безусловно, сочетаются, дополняя друг друга.

В большинстве работ по построению ИПС постулируется наличие готовой онтологии или онтологий. Отметим, что автоматизация построения онтологий является открытой и недостаточно развитой, что определяет актуальность исследований.

А/М *Глибовець, А.М. Глибовець, А.С. Шабінський*

ЗАСТОСУВАННЯ ОНТОЛОГІЙ ТА МЕТОДІВ ТЕКСТОВОГО АНАЛІЗУ ПРИ СТВОРЕННІ ІНТЕЛЕКТУАЛЬНИХ ПОШУКОВИХ СИСТЕМ

Розглянуто основні методи текстового аналізу та базових ролей онтологій при створенні інтелектуальних пошукових систем інформації. Наведено приклади

використання онтологій на етапах збору та класифікації інформації, інтерпретації користувацьких запитів, безпосередньо пошуку релевантної інформації в документі.

N.N. Glybovets, A.N. Glybovets, A.S. Shabinskiy

THE ONTOLOGIES AND METHODS OF TEXT ANALYSIS APPLICATION IN DEVELOPMENT OF INTELLIGENT SEARCH SYSTEMS

The review and detailed analysis of the main text analysis methods and basic roles of the ontologies in development of intelligent search systems are made. There are examples of using ontologies on different stages of such systems.

1. *Manning C.D., Raghavan P., Schütze H.* Introduction to information retrieval. — New York : Cambridge University Press, 2008. — 496 p.
2. *Gruber T.R.* Toward principles for the design of ontologies used for knowledge sharing // Intern. J. Human-Comput. Stud. — 1993. — **43**. — P. 907–928.
3. *Мейтус В.Ю.* Інтелектуальні системи, онтології та онтологічні системи // Наукові записки НАУКМА. Комп'ютерні науки. — 2009. — **99**. — С. 4–14. — http://www.nbuv.gov.ua/portal/soc_gum/naukma/Comp/2009_99/_01_mejtus_vyu.pdf.
4. *Salton G., Wong A., Yang C.S.* A vector space model for automatic indexing // Com. of the ACM. — 1975. — **18**, N 11. — P. 3–14.
5. *Becker J., Kuropka D.* Topic-based vector space model // Business Inform. Systems, Proc. of BIS 2003. Colorado Springs, USA. — www.kuropka.net/files/TVSM.pdf.
6. *Polyvyanyy A.* Evaluation of a novel information retrieval Model: eTVSM. — Potsdam : HPI, 2007. — bpt.hpi.uni-potsdam.de/pub/.../Thesis_Artem_Polyvyanyy.pdf.
7. *Kuropka D.* Modelle zur repräsentation natürlichsprachlicher dokumente. — www.kuropka.net/pubs.shtml.
8. *Espinasse B.F.S., Freitas F.* Agent and ontology based information gathering on restricted web domains with AGATHE. Fortaleza, 2008. SAC'08. — alarcos.inf-cr.uclm.es/Alarnet/.../2008-SAC-Reynoso.pdf.
9. *Oates T., Nagendra Prasad M.V., Lesser V.R.* Cooperative information gathering : A distributed problem solving approach // IEE Proc. on Software Engineer. — 1997. — **1**, N 1. — P. 72–88. — ftp://dis.cs.umass.edu/pub/oates_IEE_97.pdf.
10. *Tran T., Cimiano Ph., Rudolph S., Studer R.* Ontology-based interpretation of keywords for semantic search // The 6th Intern. Semantic Web Conf. Proc. of the 6th Intern. Semantic Web Conf. (ISWC'07). Korea, ISWC 2007. — P. 523–536. — www.aifb.kit.edu/web/Duc_Thanh_Tran/.../en.

Получено 11.04.2011

Статья представлена к публикации членом редколлегии Ю.Г. Кривоносом.