Ю.В. Малинина

АВТОМАТИЧЕСКОЕ ВЫЯВЛЕНИЕ ТЕМАТИЧЕСКОЙ КАРТЫ ДОКУМЕНТА

ВВЕДЕНИЕ

Рост массивов полнотекстовых документов, публикуемых в электронном виде, требует новых средств организации доступа к информации. И одной из фундаментальных проблем в этой области является то, что иногда пользователь не знает точно, какую именно информацию ему хотелось бы получить, имея лишь общее представление о границах своих интересов. Это особенно актуально для научных публикаций, так как в конечном счете научное знание воплощается в тексты и познается через тексты. Избыток научной информации в текстовой форме имеет преимущества и проблемы.

1. С одной стороны, можно допустить, что абстрактная модель научного знания является производной от множества реальных текстов.
2. С другой стороны, с каждым годом становится все очевиднее, что потребность в компактификации текстовой информации только увеличивается. По мере накопления знаний мы вынуждены передавать информацию во все более сжатом виде, сжимать имеющиеся знания все сильнее и сильнее.

На сегодня существует несколько подходов к сжатию информации.

1. Классификация (classification). Задача заключается в отнесении документа к одной из нескольких заранее определенных категорий, основываясь на содержании документа. Уже с середины XX в. для упрощения поиска материала по определенной тематике используются периодические реферативные издания, названия и тематическая направленность рубрик которых изменялась в соответствии с тенденциями развития отдельных областей науки. Это наглядный пример представления информации в сжатой форме на основе классификации.
2. Кластеризация (clustering) отличается от классификации тем, что мы заранее не знаем, какие существуют категории в рассматриваемой области, или список этих категорий неполон. Документы объединяются в подмножества динамически на основе выбранных критериев. Примером этого подхода может служить индекс цити-

рования “Science Citation Index” (SCI), ориентированный на поиск новых научных публикаций в мировой системе периодических изданий по системе научных ссылок. В этом случае для кластеризации документов используется естественная исторически сложившаяся система группировки научных работ по ссылкам автора на работы его предшественников в определенной тематической области.

3. Онтологии являются эффективным средством представления и систематизации знаний и еще одним подходом к сжатию информации. Онтологии используются для формальной спецификации понятий и отношений, которые характеризуют определенную предметную область и фактически представляют экстракт знаний об определенной области. В свою очередь, тематические карты (XTM) и RDF/OWL форматы являются наиболее важными стандартами для представления знаний в онтологиях.

ТЕМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТА

Все обозначенные выше идеи сжатия текстовой информации направлены на выявление того, что стоит за словами в документе, т.е. на определение смысла и темы текста. Считается, что тема позволит представить пользователю информацию в сжатом и облегчающем понимание сути документа виде. В то же время основная трудность заключается в идентификации темы документа. Нахождение механизмов автоматического определения основной темы и подтем документа могло бы значительно улучшить эффективность классификации, кластеризации, извлечения и поиска информации.

ТЕМА ДОКУМЕНТА

Идентификация темы часто рассматривается как высокоразвитая техника извлечения метаинформации из текста документов. Тем не менее, перед рассмотрением процесса идентификации темы необходимо определить само понятие темы документа. Согласно Брауну и Юлу [1], обычно под темой документа понимается интуитивно удовлетворительный способ описания основного принципа, который связывает различные смыслы (дискурсы) в рамках его текста. Многие исследователи заняты поиском кратких идентификаторов содержания, которые бы эффективно представляли первичные

смыслы документа. На сегодня приходится констатировать факт, что под темой документа понимается проблема или ряд проблем, которым посвящен документ, т.е. то, о чем идет речь (в самом общем смысле), и определяется это интуитивно.

В нашем случае зафиксируем формулировку основной темы текста как некоторую совокупность слов, наиболее значимых для передачи смысла текста.

ТЕМАТИЧЕСКИЕ КАРТЫ

Тематические карты являются (синтаксически) стандартизированной формой семантических сетей. В концепции тематических карт больше внимания уделяется навигации между темами, чем на связь между ними. В отличие от обычных семантических сетей тематические карты предоставляют дополнительную возможность навигации и поиска.

Наглядно тематическая карта может быть представлена в виде ориентированного графа, состоящего из вершин типа «тема» (topic), соединённых рёбрами типа «ассоциация» (association). Также имеется множество «информационных ресурсов» (occurrences). Некоторые «темы» ссылаются на нужные им «информационные ресурсы». Таким образом, «информационные ресурсы» отделяются от графа «тем» и «ассоциаций», который представляет собой только каталог информации.

Происхождение тематических карт можно проследить еще в начале 1990-х годов, когда группа Дэвенпорт (Davenport group) обсуждала пути, которые позволили бы производить обмен цифровых документов. В последние годы развитие технологии тематических карт сделало ее пригодной для включения в области управления знаниями. Сегодня тематические карты (Topic Maps) – активно развиваемая XML-технология моделирования знаний, задача которой – сделать опубликованную в Web информацию более доступной людям и компьютерам. XTM – открытый стандарт представления тематических карт в формате XML.

Формат тематических карт стандартизирован в ИСО 13250:2003 (ISO/IEC 13250:2003) – http://www.topicmaps.org.

Концепция тематических карт имеет много преимуществ и включает следующее [9].

1. Тематическая карта представляет собой структуру, независимую от местонахождения ссылок, т.е. может ссылаться на любой документ.

Поэтому тематическая карта может быть использована для навигации по нескольким ресурсам.

2. По сравнению с тезаурусами нет жестко определяемых типов связей, например, ассоциативных, иерархических и эквивалентности. Таким образом, возможно создать любой желаемый тип связи. Прирост числа ассоциативных связей позволяет описывать более сложные отношения, но в то же время такую развитую сеть связей труднее создать. Следует подчеркнуть, что связи должны создаваться очень точным и согласованным способом.
3. В тематической карте можно представить несколько точек зрения по любому вопросу. Это стало возможным благодаря характеристике темы «scope», задающей контекст, в котором имя или местоположение присваивается данной теме, и контекст, в котором темы ассоциативно связаны. Эта характеристика тематической карты позволяет поддерживать объективность в коллекциях документов.
4. Тематические карты легко визуализируются. Эта функция обеспечивается приложениями тематических карт.
5. Благодаря Public Subject Identifiers, которые идентифицируют (определяют) предметные темы, несколько тематических карт могут объединяться в одну карту.
6. Механизм, называемый веб-сервисом тематических карт, позволяет обмениваться фрагментами тематических карт. Он может быть использован для создания сетевой кооперации веб-приложений .
7. Прослеживается большая связь между тематическими картами и RDF, следовательно, проекты, основанные на тематических картах, могут быть гармонично включены в семантический веб.

ПРЕДЛАГАЕМЫЙ ПОДХОД

Идентификация темы документа

Для построения тематической карты документа необходимо идентифицировать темы и подтемы документа. Эта задача требует определения синтаксической структуры, а также семантики текста. На практике это означает глубокий анализ синтаксиса и семантики, что является трудной задачей из-за сложностей естественного языка [6]. С другой стороны, поскольку мы заинтересованы только в семантических связях, которые являются особенностями содержания документа, то определенные языковые закономерно-

сти, обнаруженные в документах, ограничены. Подходы к построению семантических сетей и выявление терминов для научных текстов были более подробно рассмотрены в предыдущих работах [7, 8].

Рассмотренная ниже технология тематического анализа позволяет дополнительно автоматически выявлять ключевые темы текста. В ходе тематического анализа устанавливаются ассоциативные связи между темами. Совокупность тем со связями образует ассоциативную семантическую сеть, представляющую модель текста, которая в дальнейшем представляется в виде тематической карты текста или совокупности текстов

Связанность текста (лексические цепочки)

Для любого данного документа слова или фразы должны быть всегда лексически связаны в цепочки вокруг центральной темы. Понятие лексической цепочки вытекает из работы, связанной с понятием текстовой сплоченности в лингвистике [4]. Согласно этой теории, смысл документа заключается в том, чтобы быть связанным и удерживать эту связь сквозь всю структуру текста через грамматическое единство, которое достигается с помощью таких средств языка, как ссылки, замены и семантически связанные слова. Лексические сплоченности неявно существуют не только между парами терминов, но и между последовательностями слов. Для последнего случая используется понятие лексической цепочки, впервые предложенное Morris J. и Hirst G. [5].

Если рассмотреть текст, то можно заметить, что слова и словосочетания, близкие по смыслу к словам основной темы, образуют лексические цепочки, которые пронизывают весь текст. Естественно предположить, что, если имеется лингвистический ресурс, в котором описаны разнообразные смысловые связи между словами (WordNet, RusNet), то можно двигаться по тексту, находить связанные по смыслу слова и формировать лексические цепочки [2]. Самые частотные (или выделенные по другим критериям) цепочки могли бы показать, чему именно посвящен конкретный текст [3].

Измерение семантической близости

Традиционный метод расчета сходства рассматривает частоты слов, но не принимает во внимание семантические отношений между словами. Подобный подход может привести в результате к неточной кластеризации. Вместо того, чтобы использовать TF и TFIDF метрики сходства, предлагается измерить семантическое сходство терминов при помощи лексической цепи. Этот метод был предложен Barzilay и Elhadad [3] и он анализирует

смысл не-терминов, итеративно измеряя семантическое сходство между парами терминов и группируя термины в кластеры, которые в конечном итоге представляют концепции каждого документа или набора документов.

Нахождение лексических цепочек

Лексическая цепь, согласно [10], представляет собой набор семантически связанных слов в тексте. На этом шаге мы ищем семантику слов в WordNet (RusNet) и вычисляем вес сходства в соответствии с отношением гипонимии в WordNet (RusNet). Шаги построения лексической цепи в основном следуют алгоритму, предложенному в работах [3, 11]:

Первоначально слово-кандидат выбирается из общих слов документа, которые не являются фразами [3]. Затем эти слова сравниваются друг с другом в соответствии с порядком слов в этом документе. Кроме того, для слова-кандидата смысл раскрывается на основе WordNet (RusNet) словаря [11].

На этом этапе рассматриваются все смыслы слова, и каждый смысл сохраняется на разных уровнях. Первый уровень представляет собой набор синонимов и антонимов, а второй – набор первых гипонимов/гиперонимов, а также их вариаций (т. е. меронимы/холонимы и т.д.).

Предположим следующую ситуацию: извлечены три слова, A, B и C; их смысловой набор, соответственно, senseA, senseB и senseC. Если для первого слова успешно найдено соответствие со вторым словом B, то они производят два вида перестановок. Первый вид перестановки – это комбинация senseA ∪ senseB, которая является первой цепочкой chain1. Вторая комбинация senseB ∪ senseC является второй цепочкой chain2. Затем сопоставляем третье слово C с A и B. Когда они совпадают успешно, они производят четыре вида перестановок, таких как chain1 ∪ senseC и chain2 ∪ senseC. При помощи этого способа нужно найти соответствие для каждого слова в документе. Если они явно совпадают, они будут производить другие лексические цепочки.

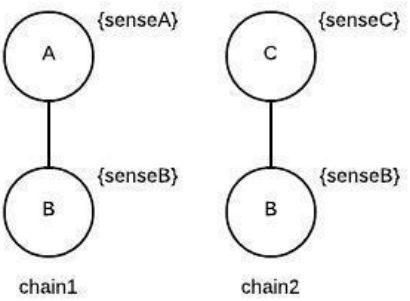

Рис. 1

Пока этот метод не вычисляет вес связи. Если обнаруживается подходящая лексическая цепь, алгоритм просто подсоединит слово к лексической цепочке. Если связь не определена, то будет строиться новая цепочка. Эта процедура будет работать итеративно, пока все соответствия не будут найдены.

Когда обнаружение всех лексических цепей закончится, в качестве следующего шага необходимо будет проверить идентичность смыслов найденных лексических цепей. На этом этапе некоторые из них могут быть объединены, поскольку они являются однородными по семантике.

Далее мы просто суммируем все связи для каждой лексической цепи, чтобы указать вес каждой. Когда лексические цепочки приобретают вес, становится возможным выбирать старшую по весу лексическую цепь и считать, что ее слова представляют основную концепцию (тему) документа, а остальные являются вспомогательными темами

ПОСТРОЕНИЕ ТЕМАТИЧЕСКОЙ КАРТЫ

Для того чтобы привязать полученную семантическую модель к интересуемой предметной области, используем словарь соответствующей тематики. В нашем случае обратимся к онтологиям, описанным в работах Князевой М.А. [12,13]. В итоговой онтологии фиксируются только те семантические конструкции, в которых участвуют термины из словаря предметной области. Далее, воспользовавшись тем фактом, что тематические карты являются (синтаксически) стандартизированной формой семантических сетей, и сходством структуры семантической сети и тематической карты,

мы можем построить относительно просто отображение узлов семантической сети в темы карты, а ссылки рассматривать как ассоциации.

На приведенной схеме изображен пример отрывка полученной тематической карты, построенной на основе множества публикаций об удалении неиспользуемого кода.

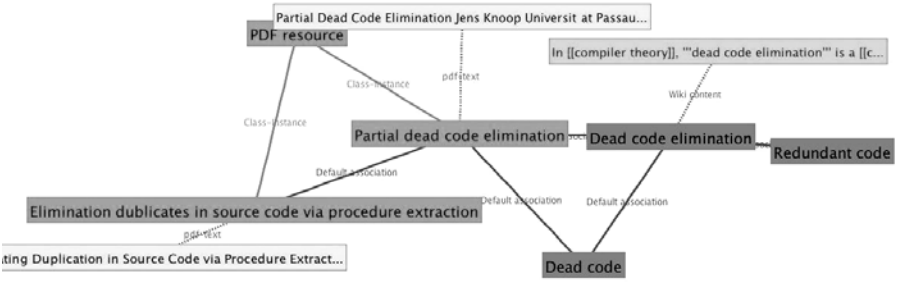

Рис. 2

СПИСОК ЛИТЕРАТУРЫ

1. Brown, G., Yule G. Discourse Analysis // Cambridge Textbooks in Linguistics Series. – Cambridge University Press, 1983.
2. Morris J., Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of a text // Computational Linguistics. – 1991. – Vol. 17 (1). – P. 21–48.
3. Barzilay R., Elhadad M. Using Lexical Chains for Text Summarization // ACL/EACL Workshop Intelligent Scalable Text Summarization. – Madrid, 1997.
4. Halliday M., Hasan R. Cohesion in English. – London: Longman, 1976.
5. Morris J., Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text // Computational Linguistics. – 1991. – Vol. 17 (1). – P. 21–48.
6. Glasgow B., Mandell A., Binney D., Ghemri L., Fisher D. An information-extraction approach to the analysis of free-form text in life insurance applications // AI Magazine. – 1998. – Vol. 19. – P. 59–71.
7. Малинина Ю.В. Семантическая сеть как формальный метод описания и обработки текстов по преобразованиям программ // Методы и инструменты конструирования и оптимизации программ. – Новосибирск, ИСИ СО РАН, 2005. – С. 137–144.

8. Малинина Ю.В. Автоматическое выявление таксономии в области преобразований программ на основе анализа семантических связей в публикациях // Методы и инструменты конструирования и оптимизации программ. – Новосибирск, ИСИ СО РАН, 2008.
9. Włodarczyk B. Topic map library = better library: an introduction to the project of the National Library of Poland // IFLA World Library and Information Congress 78th IFLA General Conference and Assembly 11–17 August 2012, Helsinki, Finland.
10. Li S.C., Wang H.C. Document Topic Detection Based On Semantic Feature // The 18th International Conference on Information Management, 2007. – P 56–73.
11. Chen K.J., Chen C. J.,Automatic Semantic Classification for Chinese Unknown Compound Nouns // Coling 2000: Proc. – P. 173–179.
12. Артемьева И.Л. Модель онтологии предметной области «оптимизация последовательных программ». Термины для описания объекта оптимизации. / И.Л. Артемьева, М.А. Князева, О.А. Купневич. – Владивосток. – 43 с. – (Преп. / НТИ 29-2000).
13. Князева М.А., Гужавин В.Д. Онтология низкоуровневой оптимизации программ // Научная сессия МИФИ-2002.