электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 005

Автоматическая классификация текстовых документов с помощью онтологий

 # 09, сентябрь 2012

Грушин М.А.

Научный руководитель: д.т.н., профессор, Норенков И.П.
МГТУ им. Н.Э.Баумана, Москва, Россия

МГТУ им. Н.Э. Баумана
bauman@bmstu.ru

Введение

Темпы роста количества документов научно-технической области в электронном виде и их доступность в сети Интернет приводят к тому, что на сегодняшний день подавляющая часть информации хранится на компьютерах в виде электронных текстовых документов. В большинстве организаций значительная часть полезных знаний содержится в документальных базах данных. Такая ситуация обуславливает повышенный интерес к области Text mining – методам автоматического извлечения и обработки знаний из текстовых документов. Получение знаний в автоматическом режиме затрудняется слабой структурированностью текстов на естественном языке. Такие знания могут быть с лёгкостью извлечены экспертом, но, с учётом огромного количества электронных документов, их эффективная обработка человеком становится весьма затратной как по времени, так и по ресурсам.

Извлечение знаний имеет своей конечной целью информационную поддержку эксперта или автоматизированной системы при принятии проектных решений. В документах, созданных специалистами, могут быть описаны подходы к решению различных проблем, рекомендации к подбору параметров и прочие знания, полезные в различных областях деятельности организации. Таким образом, основной функциональностью систем извлечения знаний является информационный поиск полезных сведений в документальных базах. Однако наряду с этой задачей должны решаться и промежуточные задачи автоматической классификации, кластеризации и аннотирования документов. В течение нескольких десятков лет разрабатываются новые методы решения подобных задач над текстами на естественном языке, а также повышается эффективность уже существующих подходов.

В системах управления знаниями (Knowledge Management) многих организаций экспертами составляются предметные онтологии, описывающие предметные области, в которых специализируются организации. Модели знаний, описываемых онтологиями, представляют собой набор ключевых понятий («концептов») и связей между ними. Применение онтологий для решения задач Text mining способствует повышению эффективности решения задач извлечения и обработки знаний из текстовых документов.

В данной работе рассматривается подход к решению задачи классификации текстовых документов с использованием ролевой кластеризации концептов онтологий, предложенной в работе [1].

Онтологии и ролевая кластеризация концептов

Онтология – один из способов представления знаний в интеллектуальных системах. Под онтологией понимают систему понятий (концептов, сущностей), отношений между ними и операций над ними в рассматриваемой предметной области, иными словами, онтология – это спецификация содержания предметной области. Например, онтология «Интеллектуальные системы» может выглядеть следующим образом:

Интеллектуальные системы = {интеллект; нейрон; нейронная сеть; обратная связь; логика; знания;…},

и также включать в себя связи между концептами вроде «*Обратная связь* – свойство *нейронной сети*».

Ролевая кластеризация концептов онтологии по классам (ролям) «объект», «средство», «свойство» и «действие» подразумевает собой распределение концептов в различные смысловые категории. Концепты этих категорий называются «простыми». Возможные комбинации (паттерны) ролей можно использовать для создания «сложных» концептов из набора простых. Таким образом, словосочетание «метод анализа быстродействия компьютера» можно рассматривать как сложный концепт, состоящий из 4-ёх простых, принадлежащих вышеприведённым ролям. В данном примере «метод» - средство, «анализ» - действие, «быстродействие» - свойство, «компьютер» - объект.

Применение подобного подхода в задачах информационного поиска по документам в пределах, по крайней мере, одной предметной области позволяет в некоторой степени учесть семантику понятий, составляющих запрос, тем самым повысив точность поиска. При запросе «анализ метода» поисковая система будет иметь представление о том, что нас интересует метод, как объект, и анализ, как действие над ним. При этом будут отсечены ненужные документы, релевантные запросу «метод анализа», которые составят практически половину выдачи системы в случае использования классических подходов к поиску. Также применение ролевой кластеризации способно повысить полноту поиска засчёт расширения запроса синонимами концептов, имеющими ту же ролевую принадлежность.

Также онтологии могут быть успешно применены при решении задач классификации и кластеризации документов, их аннотирования, упорядочения и задач поддержки принятия решений, как это описано в работе [2]. Во всех указанных случаях применения онтологии представляют собой набор значимых концептов, определяющих предметную область. Их применение позволяет избежать потерь машинного времени на анализ понятий, не входящих в предметную область, а в случае задачи классификации – не проводить временами весьма затратное обучение классификатора документов на обучающей выборке, так как классификатор представлен составленной онтологией. С другой стороны, качество решения указанных задач становится весьма зависимым от качества и полноты составленной онтологии.

Классификация документов

Для специалиста, пользующегося документальной базой данных в ходе работы, может быть интересен не весь набор документов, а только соответствующий интересующей его предметной области. Таким образом, становится актуальной задача классификации документов в базах данных по категориям. Классификация документов также имеет место в задачах фильтрации спама, распределении писем по тематике, системах электронной коммерции и многих других областях применения интеллектуальных систем. Также предварительное отнесение документа к классу в задаче информационного поиска позволяет отсечь документы, не относящиеся к тематике запроса, тем самым экономя время и вычислительные ресурсы.

Первоначальным методом классификации было ручное распределение документов по тематикам. Однако на сегодняшний день количество доступных для обработки документов огромно, а это влечёт за собой несоизмеримые с выгодой затраты средств и

времени при работе экспертов. Поэтому с 1960-х годов всё больший интерес вызывает вопрос автоматической классификации текстовых документов. Развитие методов решения задачи автоматической классификации и сами методы достаточно полно писаны в работе [3].

Изначальный подход для автоматизации работы эксперта в этой области состоял в написании им правил вида «*если - то*» для системы обработки текста, которая бы относила документ к определённой тематике при выполнении условий, заданных экспертом. Более формально классифицирующее выражение выглядит следующим образом

$$Если(ДНФ) \rightarrow То(категория),$$

где *ДНФ* – условия, выраженные в дизъюнктивной нормальной форме, *категория* – тематика, к которой следует отнести документ при истинности ДНФ. Метод прост и достаточно эффективен, однако очевидно, что он требует работы эксперта для поддержания актуальности правил и их написания.

В начале 90-х годов прошлого века экспертные правила были вытеснены методами машинного обучения. Преимущества такого подхода очевидны: системы требуют гораздо меньше экспертной поддержки и не нуждаются в написании правил классификации. Правила формируются самими системами на основе обучающей выборки. В настоящий момент наиболее популярными для решения задачи классификации являются наивный байесовский классификатор, метод Роккио, метод «*k* ближайших соседей», метод опорных векторов и различные модификации этих методов. Все эти подходы, кроме вероятностного байесовского классификатора, используют векторное представление документа, в котором содержимое представляется в виде вектора терминов, входящих в документ. Классификатор представляет собой особый документ, вектор которого формируется на этапе обучения и состоит из усреднённых значений весов терминов, входящих в документы обучающей выборки. Указанные методы имеют довольно много общего и отличаются лишь методом обучения и составления вектора-классификатора. Сама классификация является вычислением угла между двумя векторами, как степени их схожести: если вектор документа близок к вектору классификатора, то документ будет отнесён к данной категории.

Если для классификации используется онтология предметной области, то вектор документа можно сравнивать с вектором самой онтологии. Отсюда следует два важных отличия от классических методов машинного обучения. Применение онтологий позволяет отказаться от этапа обучения классификатора. Описание предметной области в виде онтологии само является классификатором, таким образом, не тратится время и вычислительные ресурсы на построение среднего документа из обучающей выборки. Второе отличие заключается в том, что при таком подходе в вектор документа включаются только те термины, которые включены в рассматриваемую онтологию. Это значит, что те понятия, которые не входят в набор концептов онтологии, уходят из процесса вычисления весов терминов. Также имеется отличие классификатора в виде онтологии от классификатора в виде усреднённого документа. В обоих случаях классификатор является моделью «эталонного» документа, соответствующего предметной области. Но если он является «усреднённым» документом, то в его состав могли войти термины, употреблявшиеся в документах, но не имеющие отношения к описываемому разделу. В случае онтологии, напротив, классификатор является описанием предметной области без каких-либо лишних для неё понятий. В таком случае этот классификатор является более универсальным с точки зрения использования его в составе различных систем и для различных задач.

Степень соответствия документа классу (онтологии) рассчитывается, как сумма весов всех терминов данной онтологии, найденных в документе:

$$R_{dC} = \sum_{t \in C} w_{td},$$

где R_{dC} - степень соответствия документа *d* кластеру *C*, w_{td} - вес термина *t* в документе *d*.

Существуют различные подходы к взвешиванию терминов в документе, о них можно узнать в источнике [4]. С учётом того, что концепты в онтологиях разделены по ролям, необходимо ввести различные веса для ролей, а также отдельно обрабатывать сложные концепты. Хорошо зарекомендовал себя следующий способ взвешивания

$$w_{td} = \begin{cases} tf, \text{если концепт простой и его роль "объект"}, \\ 0.1 \cdot tf, \text{если концепт простой}, \\ (1+k) * tf, \text{если концепт сложный}, \end{cases} \quad (1)$$

где *tf* – количество вхождений концепта в документ, *k* – коэффициент, учитывающий сложность концепта.

Следует отметить неравномерность значимости разных ролей концептов для классификации документа. Существует предположение, что предметная область наиболее полно характеризуется концептами типа «объект», в то время как концепты остальных типов могут быть инвариантны по отношению к разным предметным областям. Это было подтверждено экспериментально в ходе проверки классификатора – если присваивать всем простым концептам, встреченным в документе, равные веса, соответствующие количеству их вхождений в текст, то документ будет с равной вероятностью отнесён ко всем классам сразу. Отсюда следует, что наиболее значимыми нужно сделать понятия-объекты, а остальным присваивать веса по другой схеме. Как видно из формулы (1), предложенная схема взвешивания учитывает это замечание. Также в формуле (1) учитывается повышение веса сложного термина в зависимости от количества простых, составляющих его.

В серии экспериментов классификатор с использованием онтологий показал себя хорошо, процент правильно классифицированных документов довольно высок. Остаётся лишь отметить, что в данном подходе качество классификации напрямую зависит от качества созданной онтологии.

Заключение

В настоящее время всё больше организаций используют в своей работе системы управления знаниями, которые используются для решения широкого круга задач. В данной работе рассмотрен подход к решению одной из таких задач – классификации документов. С одной стороны этот подход имеет тот же минус, что и подход с составлением правил классификации вручную – требуется работа экспертов для создания описаний предметных областей и поддержания их актуальности. Но с другой стороны такой подход является более гибким. Онтологии могут быть использованы в ряде других задач управления знаниями, поэтому их создание и поддержка более оправдано с точки зрения затрат. Также отнесение документа к тематике происходит не просто по факту наличия в нём определённых терминов, но на основе вычисления меры близости документа и онтологии. Таким образом, рассмотренный подход в некоторой степени объединяет в себе эффективность старого и универсальность современного подхода к классификации документов.

Литература

1. Норенков И.П., Уваров М.Ю. Поддержка принятия решений на основе паттернов проектирования. // Электронное научно-техническое издание «Наука и образование», 2011, 9.
2. Норенков И.П. Задачи управления знаниями, извлекаемыми из текстовых документов. // Электронное научно-техническое издание «Наука и образование», 2011, 9.

3. Fabrizio S. Machine learning in automated text categorization// ACM Computing Surveys, 2002, 34(1), pp 1-47.
4. Manning C., Raghavan P., Schütze H.. Introduction to Information Retrieval. - Cambridge University Press, 2008, 544 p.
5. Bevainyte A., Butenas L. Document classification using weighted ontology// Materials Physics and Mechanics, 2010, №9, pp. 246-250.