УДК 681.325.65

Ха Ти Чунг

РАЗРАБОТКА МОДЕЛЕЙ ПРЕДСТАВЛЕНИЯ ДОКУМЕНТОВ И КЛАССИФИКАТОРА НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ

Введение. В последнее время колоссально возрастает объем цифровых изданий и документов в хранилищах разных информационных систем (в локальных сетях, в электронных библиотеках, электронных каталогах и т.д.). Актуальным стал вопрос организации и поиска документов в этих системах. Ограниченные возможности инструментариев для навигации и поиска существенно сокращают возможность получать наиболее актуальную и полную информацию по конкретной тематике. Как следствие, это делает немалую часть данных бесполезной, либо возвращает неполное представление о проблеме, поскольку для получения нужной информации требуется большие трудозатраты пользователей на непосредственный анализ информации по интересующей теме.

Автоматизированная рубрикация (классификация) документов является одним из способов повышения эффективности поиска и доступа к информации. Под данной задачей понимается автоматическое отнесение (физически и/или логически) документов к одной или нескольким тематикам (рубрикам) из конечного множества рубрик. Данная задача остается до сих пор актуальной задачей в области информационного поиска, в которой многие проблемы далеки еще от окончательного решения. Помимо общих проблем для всех задач классификации, к данной задаче дополняется проблема определения семантических признаков естественного языка. Более того, при решении реальной задачи, например, при классификации документов различных форматов (условно разделены на группы: графических – отсканированных документов, гипертекстовых, офисных документов и др.) возникают и другие проблемы. Например, такие как проблемы работы с логической организацией и форматом представления документов.

В работе [2] показывается, как с помощью имеющихся средств автоматического преобразования форматов можно привести эти документы полностью или частично к текстовому формату. Таким образом, имеется возможность применения методов текстовой классификации для решения задачи рубрикации документов. С этой целью для решения задачи рубрикации документов в данной работе описывается модель представления документов различных форматов, также формулируется модель классификатора на основе нечеткой логики. Модели основываются на гипотезе о том, что встречающиеся информационные единицы (слова, словосочетания, другие объекты (далее термы)), и характеристики употребления этих единиц зависят от тематики документа [3].

Краткий обзор моделей представления документов. Первые модели представления документов были разработаны в конце 1950-х годов. В то время дорогостоящая память и низкое быстродействие компьютеров не позволяли обрабатывать большие объемы информации. Удобно представить документ в виде множества слов. Данная модель исторически получила название модель “*bag of words*”. С развитием технологических и аппаратных средств были разработаны более сложные модели на основе классической модели “*bag of words*”. Многие исследователи пытались включить в эти модели как характеристики термов и отношений между ними, так и другие признаки естественного языка. Наиболее распространенные модели, успешно применены в реальных задачах классификации и состоят в использо-

вании описания текста как нормированного вектора термов (обычно просто слов) в евклидовом пространстве термов. Более сложные модели, разработанные за последнее время, учитывали, в той или иной степени, различные признаки документов, а также спецификации конкретных задач. К этим признакам можно отнести:

1. Слова, находящиеся в определенных полях по стандартизованным шаблонам.
2. Статистика слов в тексте документа.
3. Взаимное положение слов в тексте, связанность и последовательность слов текста.
4. Оформление и местоположение слов.
5. Парадигматическое, ассоциативное отношение между словами (такие как отношение синонимии, антонимии, полисемии, омонимии, отношение род-вид, и др.).
6. Взаимные связности между документами по разным критериям.

Формулировка задачи рубрикации документов. Пусть на некотором этапе для интересующей предметной области знания было создано конечное множество тематик (рубрик) и требуется согласно этим тематикам расклассифицировать документы. Эту задачу решает текстовый классификатор.

Введем следующие определения.

Моделью текстового классификатора Z называется пятерка:

$$Z = (D, C, H, I, \Phi), \quad (1)$$

где $D = \{d_1, \ldots, d_{|D|}\}$ – множество документов (не обязательно конечное); $C = \{c_1, \ldots, c_N\}$ – множество классов-рубрик, где N – число классов; $H = \{(c_i, c_j),\ i, j \in 1 \ldots N\}$ – иерархия классов–рубрик. В паре (c_i, c_j) рубрика c_i является родительской (более общей) по отношению к рубрике c_j; $I = \{I(c_i),\ \ i = 1 \ldots N\}$ – множество описаний (образов) классов. Описание каждого класса $I(c_i)$ представляет собой множество признаков рубрик c_i. Данное множество может формироваться автоматически в процессе обучения и/или задаваться экспертным путём; $\Phi = \{\mu_{ji}\}$ – функция рубрикации. Здесь μ_{ji} неотрицательное число, приписанное каждой паре $(d_j, c_i) \in D \times C$ и обозначающее степень принадлежности документа d_j классу c_i, Ограничим это число интервалом $0 \le \mu_{ji} \le 1$.

Обычно для построения функции рубрикации используют метод обучения классификатора на некоторой выборке документов $D_{teach} = \{d_1, \ldots, d_{|D_{teach}|}\} \in D$. При этом в обучающей выборке заранее известно μ_{ji} значение для каждой пары $(d_j, c_i) \in D_{teach} \times C$. В обучающую выборку для каждой тематики собирают только такие документы, для которых $\mu_{ji} \ge \mu_0^i$, где μ_0^i – некоторое наперед заданное пороговое число, которое определяет, можно ли использовать документ в качестве обучающего для класса c_i. После обучения классификатор считается построенным, и он применяется ко всему множеству документов D.

Разработка модели представления документа. Модель представления документа должна позволять: уменьшить потери информации (что позволяет повысить точность классификации); гибко описать документы различных форматов;

сократить трудозатраты и время разработки на реализацию классификатора на практике.

Очевидно, что каждый документ, в независимости от вида логической организации и физического формата, можно представить в виде основного текста, который описывает его семантическую сущность и некоторую дополнительную информацию, которая определяет некоторый стандартизированный формат и способ составления и оформления такого рода документов. Например, книги кроме основного текста содержат оглавления, аннотации, список трудов и т.д. Свои требования предъявляются к оформлению журнальных статей, научно-технических отчетов и т.п.

Исходя из этого, введем следующую формальную модель документа. Пусть каждый документ представляет собой четверку:

$$d_j = \{M_j^1, M_j^2, T_j^d, T_j^u\}, \quad (2)$$

где:

1. M_j^1 – группа атрибутов внешнего описания. Эти атрибуты не всегда содержаться в тексте документа. Например, сюда можно отнести:
   a) информацию об источнике (автор, издательство, URI, URL и т.д.);
   b) информацию о времени создания и модификации документа;
   c) информацию о формате и размере документа;
   d) и т.д.

   Очень часто такие характеристики, как название документа, имя автора, рассматривается отдельно от текста документа. Также значения некоторых из атрибутов можно получить непосредственно из дескриптора файла документа.
2. M_j^2 – термы, извлеченные из разных функциональных областей документа, но не из «тела» документа (текста документа). Можно обозначить некоторые из функциональных областей: название документа, ISBN, УДК, ББК, аннотация.
3. T_j^d – термы, полученные из анализа текста документа. Разумеется, не все термы из текста документа, а только те, которые удовлетворяют некоторым условиям при анализе. Именно термам, находящимся в этих областях, нужно присвоить высокие коэффициенты. Они используются при оценке значимости термов. Такой подход основывается на гипотезе о том, что эти области концентрируют важную информацию о тематике документа.

   Множество T_j^d следует разбить на два подмножества W_j^d и O_j^d: W_j^d – подмножество термов (слов и словосочетаний); O_j^d подмножество прочих объектов (наименований, сокращений, аббревиатур, цифробуквенных комплексов и др.).
4. T_j^u – термы, вводящиеся пользователями для корректировки тематики документа.

Описание тематик. Каждое описание $I(c_i)$ для тематики c_i может представлять собой двойку:

$$I(c_i) = \{T_i^c, U_i^c\}, \quad (3)$$

где T_i^c – множество термов, приписанных к рубрике на шаге ее составления или обучения классификатора; U_i^c – множество вводящихся термов пользователями – специалистами по предметной области или полученных в результате экспертного анализа.

Подход к задаче классификации, основанный на нечеткой логике. Учитывая выше предложенную модель классификатора, для каждого терма t_k из документов можно вычислить степень его принадлежности $\mu(t_k, c_i)$ к каждой c_i из тематик, $\mu(t_k, c_i) \in [0,1]$. Функция $\mu(t_k, c_i)$ вычисляется следующим образом:

$$\mu(t_k, c_i) = \alpha_1 \cdot f(t_k, c_i) + \alpha_2 \cdot L(t_k, c_i) \cdot R(t_k, c_i), \quad (4)$$

где $f(t_k, c_i)$ – итоговая функция от статистических характеристик терма t_k (место появления в тексте – в группе M_j^2 или T_j^d, частотные характеристики, например, TIFIDF [5], длина текста документа и др.);

$L(t_k, c_i)$ итоговая функция от взаимоотношений t_k с другими термами в описании c_i. В частном случае, пусть $P(t_k \mid c_i)$ – вероятность того, что терм t_k принадлежит рубрике c_i, согласно [6] определяется на этапе обучения на основе обучающего множества D_{teach}, следующим образом:

$$P(t_k \mid c_i) = \frac{1 + \sum_{d_j \in D_{teach}} \theta(t_k, d_j) P(c_i \mid d_j)}{2 + \sum_{d_j \in D_{teach}} P(c_i \mid d_j)}, \quad (5)$$

где $\theta(t_k, d_j) = 1$, если $t_k \in d_j$, иначе $\theta(t_k, d_j) = 0$, пусть:

$$P(c_i) = \frac{|c_i|}{|D_{teach}|},$$

здесь $|c_i|$ – количество документов из обучающей выборки, приписанных рубрике c_i, пусть величина $P(t_k) = \frac{N(t_k)}{\sum_{d \in D_{teach}} |d|}$, где $N(t_k)$ – количество появления t_k в обучающем множестве.

Тогда согласно [4, 7] значение $L(t_k, c_i)$ можно считать:

$$L(t_k, c_i) = MI(t_k, c_i) = \log\left(\frac{P(t_k \mid c_i)}{P(t_k) P(c_i)}\right), \quad (6)$$

где $R(t_k, c_i)$ – авторитетность терма $t_k \in T$ для класса c_i. Экспертным путем можно задавать или определять ранг термов, которые должны присутствовать в тематике c_i; α_1, α_2 коэффициенты для регулирования влияния компонентов на результат вычисления $\alpha_1 \le 0.5, \alpha_2 \le 0.5$.

Аналогично, функцию принадлежности терма t_k документу d_j – $\mu(t_k, d_j)$ можно вычислить следующей формулой:

$$\mu(t_k, d_j) = \beta_1 \cdot f(t_k, d_j) + \beta_2 \cdot L(t_k, d_j) \cdot R(t_k, d_j). \quad (7)$$

Таким образом, каждый документ d_j и тематику c_i можно представить в виде нечетких множеств:

$$\begin{aligned} I(c_i) &= \{(\mu(t_k, c_i)/t_k)\},\ t_k \in c_i, \\ d_j &= \{(\mu(t_k, d_j)/t_k)\},\ t_k \in d_j. \end{aligned} \quad (8)$$

Для уменьшения их размерности предлагается использовать α-уровень [1] этих множеств, т.е:

$$I(c_i) = \{(\mu(t_k, c_i)/t_k)\},\ t_k \in c_i, \mu(t_k, c_i) \geq \alpha, 0 \leq \alpha < 1. \quad (9)$$

Значение μ_{ji} для каждой пары $(d_j, c_i) \in D \times C$ вычисляет следующей формулой:

$$\mu_{ji}(c_i, d_j) = \frac{\sum_{t_k \in d_j} \sum_{t_m \in c_i} \mu_{kj}^d \,\&\, \mu_{mi}^c}{|d_j| \cdot |I(c_i)|}. \quad (10)$$

Функция рубрикации Φ при этом представляет собой нечеткое соответствие:

$$\Phi = \{\mu_{ij}(c_i, d_j)/(c_i, d_j)\},\ (c_i, d_j) \in C \times D. \quad (11)$$

Из выше изложенной постановки, суть задачи заключается в определении нечеткого соответствия $\Phi = \{\mu_{ji}\}, 0 \leq \mu_{ji} \leq 1$, которое присваивается каждому из входных документов – нечеткое множество термов, некоторое числовое значение – степень принадлежности документа к рубрике.

Таким образом, в рамках предложенной модели задача рубрикации документов состоит из следующих этапов:

1. Для каждого документа определяется набор атрибутов, метаданных и термов по модели (2). Эта операция для каждого документа происходит один раз на этапе распознавания и индексирования.
2. На основе обучающей выборки документов определяются описания тематик-классов (3).
3. Определяются значения $\mu_{ji}(c_i, d_j)$ по (4-11) и строится соответствие Φ.
4. Осуществляется итерационное уточнение классификатора экспертным путем или повторным обучением.

В заключение можно сделать вывод о том, что предложенная в данной работе модель классификатора позволяет реализовать на практике систему классификации документов различных форматов. Предложенная модель представления документа позволяет гибко описать научно-электронные издания, что характерно для электронных ресурсов разных библиотечно-информационных систем. Повышение точности метода достигается на пути замены слов устойчивыми синтаксическими группами. Анализ зависимости $f(t_k, c)$ и связи $L(t_k, c)$, а также способ задания

ранговой степени $R(t_k, c)$ для каждого из выбранных термов является предметом отдельного исследования и будет освещен автором в отдельной работе.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Блюмин С.Л., Шуйкова И.А., Сараев П.В., Черпаков И.В.* Нечеткая логика: алгебраические основы и приложения: Монография. – Липецк: ЛЭГИ, 2002. – 113 с.
2. *Ха Т.Ч., Юрчук С.Н.* Создание текстовой выборки на основе электронного архива данных лаборатории ELDIC для исследования задач автоматической обработки текстов на естественном языке // Труды Всероссийской научной школы-семинар молодых ученых, аспирантов и студентов Таганрог: "Интеллектуализация информационного поиска, скантехнологии и электронные библиотеки". – Таганрог: Изд-во ТТИ ЮФУ, 2008. – С. 82-86.
3. *Igor Kuralenok, Vladimir Dobrynin, Igor Nekrestyanov, Mikhail Bessonov, and Ahmed Patel.* Distributed search in topic-oriented document collections. In Proc. of World Multiconference on Systematics, Cybernetics and Informatics (SCI'99), volume 4, pp. 377-383, Orlando, Florida, USA, August 1999.
4. *Dumais S.T., Platt J., Heckerman D., Sahami M.* Inductive learning algorithms and representations for text categorization. In Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management (Bethesda, MD, 1998), 148-155.
5. Haris Z. Mathematical Structures of Language. Interscience Publishers John Wiley & Sons, New York. – 1968, 80. – 230 p.
6. *McCallum A., Nigam K.* A Comparison of Event Models for Naive Bayes Text Classification. In AAAI/ICML-98 Workshop on Learning for Text Categorization, 1998. p. 41-48.
7. *Sebastiani F.* Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1): 1-47, 2002.

УДК 517.714.3

Г.В. Уралев

КОНЦЕПТУАЛЬНАЯ СПЕЦИФИКАЦИЯ НЕКОТОРЫХ ЗАДАЧ ИНФОРМАТИКИ*

Введение. Здесь под *задачей информатики* мы понимаем задачу, решение которой достигается с помощью программ, использующих экспертное знание. Важным этапом в разработке таких программ является *концептуальная спецификация* структуры этого знания, в частности, знания предметной области и методов решения задачи. Концептуальная спецификация состоит в идентификации соответствующих понятий предметной области, связей между ними, схем, правил, процедур и т.п.

Часто концептуальная спецификация выполняется на неформальном уровне. Но тенденция такова, что все чаще используются формальные языки для концептуальной спецификации, благодаря которым удается строить концептуальные модели приложений, допускающие формальную интерпретацию и, следовательно, машинную обработку. Процесс построения таких моделей называют *концептуальным моделированием* [1]. Формализмы концептуальной спецификации называются *концептуальными языками* или *языками концептуального моделирования* [2]).

Одним из важных требований, предъявляемых к концептуальным языкам, является их когнитивная адекватность, т.е. близость их конструкций к тем, какие использует эксперт, выполняющий неформальную концептуализацию. Так как такая

* Работа выполнена при финансовой поддержке РФФИ (грант № 08-01-00465).