ПОСТРОЕНИЕ МЕРЫ БЛИЗОСТИ ДОКУМЕНТОВ ДЛЯ ТЕМАТИЧЕСКОЙ КЛАССИФИКАЦИИ АВТОРЕФЕРАТОВ ДИССЕРТАЦИЙ

Леонова Ю.В.[(1)], Федотов А.М.[(2)], Федотова О.А.[(3)]

[(1)] Институт вычислительных технологий СО РАН, г. Новосибирск
[(2)] Новосибирский государственный университет, г. Новосибирск
[(3)] Государственная научно-техническая библиотека СО РАН, г. Новосибирск

В работе рассматривается метод тематической классификации авторефератов диссертаций. Для этого используется специально построенная мера близости документов, учитывающая специфику предметной области. В качестве шкал для определения меры предлагается брать характеристики структурных атрибутов описания авторефератов (научная новизна; положения, выносимые на защиту и т.п.). Значения весовых коэффициентов в формуле для вычисления меры близости определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы.

Ключевые слова: тематическая классификация, мера близости, предметная область, весовые коэффициенты.

CONSTRUCTION OF A MEASURE OF PROXIMITY OF DOCUMENTS FOR THE THEMATIC CLASSIFICATION OF ABSTRACTS OF DISSERTATIONS

Leonova Yu.V. [(1)], *Fedotov A.M.* [(2)], *Fedotova O.A.* [(3)]

[(1)] Institute of Computational Technologies SB RAS, .Novosibirsk, Russian Federation
[(2)] Novosibirsk State University Novosibirsk, Russian Federation
[(3)] State Public Scientific and Technical Library SB RAS, Novosibirsk, Russian Federation

The method of thematic classification of abstracts of dissertations is considered in the work. For this purpose, a specially constructed measure of the proximity of documents is used, taking into account the specifics of the subject area. As scales for determining the measure, it is suggested to take the characteristics of the structural attributes of the description of the author's abstracts (scientific novelty, provisions to be defended, etc.). The values of the weight coefficients in the formula for computing the proximity measure are determined by the assumed a posteriori reliability of the data of the corresponding scale.

Ключевые слова на английском языке: thematic classification, measure of proximity, subject area, weighting factors.

Введение. Задачи поиска и выделения информации является одной из важнейших задач, возникающих при построении информационных систем. Пользователь ищет не документы как таковые, а сокрытые в них факты или содержимое для удовлетворения собственных информационных потребностей. Универсальным подходом, решающим эту задачу, является тематическая классификация документов. К тому же, как было отмечено Дональдом Кнутом (см. [1]) задачи поиска и классификации документов являются двойственными, то нам достаточно рассмотреть модель классификации документов, наиболее адекватно отражающую особенности работы с информацией.

Наиболее распространенным вариантом классификации библиографических ресурсов является фасетная классификация, теория построения которой формализована индийским библиотековедом Ш.Р. Ранганатаном (см. [2]). Объекты классифицируются одновременно по нескольким независимым друг от друга признакам (фасетам). Применительно к цифровым документам (и электронным ресурсам вообще) в качестве фасе-

тов выступают элементы метаданных, которые включат и ключевые термины.

Кратко фасетная классификация состоит в следующем:

Определяется множество тематических классов документов. Класс имеет несколько фасетов, соответствующих различным аспектам классифицируемого понятия, разделы, в свою очередь, могут иметь подразделы и т.д.

Из коллекции изучаемых документов выписываются все существенные термины, которые группируются по фасетам, т. е. объединяются в соответствующие классы.

Термин, принадлежащий некоторому фасету, называется его фокусом. При индексировании документов их содержание выражается последовательностью фокусов.

В работе предложена формальная модель фасетной классификации, основанная на индексации документов ключевыми терминами, выбираемыми из некоторого словаря. Предложен и апробирован алгоритм классификации, основанный на специально построенной мере близости, учитывающий специфику классификационной модели. В качестве базы для экспериментов выбрана коллекция, состоящая из 4000 авторефератов. Мы остановили свой выбор на авторефератах диссертаций по следующим причинам: практически одинаковый объем и наличие структуры, позволяющее изучить иерархию фасетов.

1. Модель классификации. Простейшая формальная модель классификации документов с использованием метаданных (ключевых терминов) документов выглядит следующим образом [3, 4]. Рассмотрим коллекцию документов $D = \{d_i\}$. Любой документ d_i. из коллекции D представляется как $d_i = <m_i^{j,k}>$, где $m_i^{j,k}$ – значения элементов метаданных T^j, k – количество значений (с учетом повторений) соответствующего элемента метаданных в описании документа. Рассмотрим подмножество метаданных T_C, определяющее набор классификационных признаков документов. Для фиксированного элемента метаданных T^j, где $T^j \subset T_C$, заранее определяются подмножества T_i^j множества значений этого элемента метаданных (указанные подмножества могут, вообще говоря, пересекаться). Множество документов разбивается на классы эквивалентности, соответствующие различным значениям или же заранее выбранным подмножествам множества значений этого элемента метаданных.

Будем считать два документа толерантными, если у них совпадает значение хотя бы одного из элементов метаданных, входящих в T_C или (напомним, что толерантность – отношение, которое обладает свойствами рефлексивности и симметричности, но, вообще говоря, может не обладать, в отличие от отношения эквивалентности, свойством транзитивности). Каждое такое значение порождает класс толерантности [5].

Рассмотрим всевозможные сочетания значений элементов метаданных, входящих в T_C. Множества документов, обладающие одинаковым набором значений, суть ядра толерантности, которые служат классами эквивалентности на множестве документов. С содержательной точки зрения этой ситуации соответствует вхождение некоторого раздела классификатора в раздел более высокого уровня, когда оба этих раздела учитываются при описании пространства толерантности (разумеется, можно и не учитывать раздел более низкого уровня при определении толерантных элементов, но тогда мы будем иметь дело с пространством толерантности, отличным от первоначального).