Режим доступа: http://shcherbak.net/ avtomatnoe-predstavlenie-ontologij-i-operacii-na-ontologiyax.

3. Рабчевский Е.А. Автоматическое построение онтологий / Интернет-ресурс. – Режим доступа: http://shcherbak.net/avtomaticheskoe-postroenie-ontologij.

4. Рабчевский Е.А. Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска // Труды 11 Всероссийской науч. конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – Петрозаводск, 2009. – С. 69-77.

5. Мозжерина Е. С. Автоматическое построение онтологии по коллекции текстовых документов // Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции – Воронеж, 2011 – С. 293 – 298.

Калинин А.С., Охрименко К.С.
Науч. руководитель к.т.н., доц. Вороной С.М.
Донецкий национальный технический университет

Рубрикация текстов на основе автоматического построения семантических сетей

Задача рубрикации (классификации) документов, то есть отнесение документа к одной или нескольким темам, является весьма актуальной в связи с ростом объема доступной полнотекстовой информации. Из-за чего классифицировать вручную большие объемы информации становится практически невозможным. Поэтому с каждым днем все больше возрастает необходимость создания алгоритмического и программного обеспечения, которое позволяло бы в автоматическом режиме и с максимальной точностью классифицировать тексты различного содержания [1].

Классификация/рубрикация информации (отнесение порции информации к одной или нескольким категориям

из ограниченного множества) является традиционной задачей организации знаний и обмена информацией, рассматривается как одна из классических задач информационного поиска. Распространенность больших информационных коллекций делает необходимым развитие автоматических методов рубрикации.

Известны две основных технологии автоматической рубрикации:

– методы, основанные на знаниях (также именуемые «инженерный подход»), при применении которых правила отнесения текстов к рубрикам строятся инженерами по знаниям в форме булевских выражений, правил продукций;

– методы на основе машинного обучения, при применении которых используется коллекция документов, предварительно рубрицированная человеком; алгоритм машинного обучения строит процедуру классификации документов на основе автоматического анализа заданного множества рубрицированных текстов.

Оценка качества автоматической классификации производится путем сравнения с эталонной («правильной») классификацией набора документов, т. е. на основе коллекции документов, рубрицированных вручную.

В данной исследовательской работе предлагается новый механизм рубрикации текстов на основе построения семантических сетей. Рубрикации больших текстовых коллекций осуществляется по заранее фиксируемому множеству рубрик. Особый акцент в работе делается на поиск оптимальных путей рубрикации текстов с использованием семантических сетей.

Семантические сети – наиболее мощная математическая модель для представления знаний о предметной области (ПО), одно из важнейших направлений искусственного интеллекта. В настоящее

время в научной литературе описано множество альтернативных представлений моделей семантических сетей. Они предназначены для решения разнообразных задач в различных ПО.

В общем случае под семантической сетью понимается выражение, приведенное в формуле 1.

$$S = (O, R_1, R_2, \ldots R_к) \quad , \tag{1}$$

где O – множество объектов конкретной предметной области; R_i $(i=1,n)$ – множество отношений между объектами; i – тип отношений.

Из множества существующих методов построения семантической сети был выбран метод создания семантической сети из коллекции текстовых документов определенной предметной области [2]. Суть метода заключается в пошаговом анализе текста, который приведен на рисунке 1.

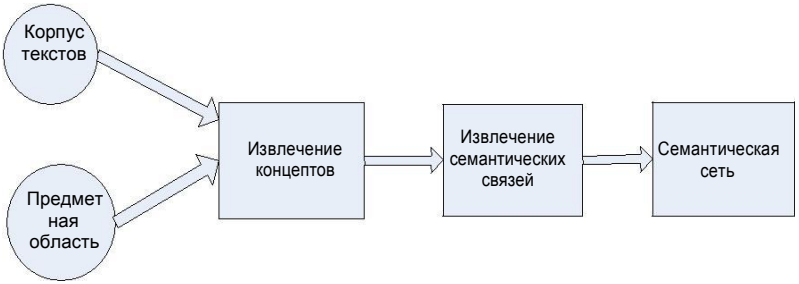

Рис. 1. Процесс создания семантической сети

На этапе извлечения концептов происходит выделение ключевых слов, выделение ключевых словосочетаний и группирование словосочетаний. В свою очередь группирование ключевых слов разбивается на несколько этапов, приведенных ниже.

1. Нормализация, токенизация, лемматизация.

2. Фильтрация на основе лингвистической информации: удаление стоп слов, имен собственных, чисел, дат, всего

остального кроме существительных и прилагательных.

3. Ранжирование слов-кандидатов с использованием статистической информации.

Выделение ключевых словосочетаний также делится на отдельные шаги: 1.Извлечение свободных словосочетаний. 2. Группирование словосочетаний-кандидатов, путем поиска наибольших общих подстрок. 3. Ранжирование словосочетаний.

После реализации всех шагов, описанных выше, получим семантическую сеть. В результате семантическая сеть будет представлять граф, состоящий из концептов и связей между ними.

Т.к. количество рубрик изначально фиксировано, то по ним можно автоматически построить эталоны в виде семантических сетей на каждую рубрику.

Используя алгоритм построения семантической сети, приведенный выше можно построить семантическую сеть классифицируемого текста.

Следующий этап, предлагаемого метода, - этап сравнения вновь построенной семантической сети текста с эталонами рубрик. Этот процесс реализуется путем пошагового сопоставления отношений и концептов новой семантической сети с эталонами и вычисления количества совпадений. Рубрика, эталон которой имеет максимальное количество совпадений отношений и концептов, является исходным классом для рубрицируемого текста.

Литература.

1. Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры [Электронный ресурс] / Б.В. Добров, Н.В. Лукашевич – Режим доступа: http://www.cir.ru/docs/ips/publications/02_cai_rubr.pdf.

2. Методы автоматической рубрикации и оценка их качества [Электронный ресурс] / Н.В. Лукашевич – Режим доступа:

http://like-money.ru/stati/259-metody-avtomaticheskoj-rubrikaczii-i-oczenka-ix-kachestva.

3. Построение семантической сети из разнородных данных [Электронный ресурс] / Панченко А. – Режим доступа: http://it-claim.ru/Persons/Panchenko/presentation 2010_sept_final.pdf.