ИСПОЛЬЗОВАНИЕ МЕР РЕЛЕВАНТНОСТИ СТРОКА-ТЕКСТ ДЛЯ АВТОМАТИЗАЦИИ РУБРИКАЦИИ НАУЧНЫХ СТАТЕЙ

Е.Л. Черняк,
аспирант кафедры анализа данных и искусственного интеллекта, отделение прикладной математики и информатики, факультет бизнес-информатики, Национальный исследовательский университет «Высшая школа экономики»

Б.Г. Миркин,
доктор технических наук, профессор кафедры анализа данных и искусственного интеллекта, отделение прикладной математики и информатики, факультет бизнес-информатики, Национальный исследовательский университет «Высшая школа экономики»

Адрес: 101000, Москва, Мясницкая ул., 20
E-mail: echernyak@hse.ru, bmirkin@hse.ru

В большинстве задач семантического анализа текстовых материалов возникает потребность в использовании мер релевантности строка-текст. К таким задачам относится и задача рубрикации научных статей. Как правило, научные статьи индексируют согласно системе рубрик, заданной таксономией – иерархической структурой рубрик (или понятий). Например, в научных журналах международной Ассоциации вычислительной техники (ACM), наиболее авторитетной в области информатики организации, статьи проиндексированы их авторами с использованием специально разработанной многоуровневой таксономии ACM CCS. В работе исследуется возможность автоматизации рубрикации научных статей с использованием мер релевантности строка-текст: в качестве строк используются темы таксономии, а в качестве текстов – непосредственно тексты научных статей или некоторые их фрагменты. Мера релевантности строка-текст ставит им в соответствие некоторое число, которое может интерпретироваться по-разному в зависимости от используемой модели релевантности. Чем больше значение показателя релевантности, тем сильнее связь между строкой и текстом.

В статье проведено экспериментальное сравнение различных мер релевантности строка-текст для автоматизации рубрикации научных статей. В эксперименте участвуют три меры: (а) косинусная мера релевантности, основанная на традиционном кодировании текстов с использованием tf-idf весов термов, (б) популярная характеристика вероятности порождения термов BM25 и (в) предложенная авторами характеристика условной вероятности символа в фрагментах, выделенных с использованием аннотированного суффиксного дерева, СУВСС. Для эксперимента использованы аннотации статей, опубликованных в журналах ACM, и таксономия ACM CCS 2012. В результате применения каждой из этих трёх мер получаются автоматические рубрикации статей – списки таксономических тем, упорядоченных по убыванию оценки релевантности данной статье. Оценка качества полученных результатов осуществляется с помощью сравнения автоматической рубрикации с авторской: чем выше в соответствующем списке авторская тема, тем точнее получившаяся рубрикация. Точность рубрикации оценивается с помощью популярных мер MAP и nDCG, а также меры, характеризующей количество вхождений авторских тем в топ списка, предложенной в данной работе. Проведённые нами эксперименты показывают, что использование СУВСС существенно повышает точность рубрикации по сравнению с другими двумя мерами релевантности.

Ключевые слова: меры релевантности строка-текст, аннотированные суффиксные деревья, рубрикация текстов, мера качества рубрикации.

1. Введение

Разработка надёжных средств автоматизации семантического анализа текстовых материалов является одной из самых насущных задач информатики. Уровень актуальности этой проблемы не может быть переоценён из-за взрывного накопления текстовых документов в интернете. В частности, внимание многих исследователей привлекает проблема категоризации или классификации текстовых документов: для заданной коллекции текстовых документов и заданного множества категорий, представленных текстовыми метками, требуется каждому документу приписать релевантные ему категории. Эта проблема является основой таких важных направлений информатики как извлечение/поиск информации [1], каталогизация документов [2], аннотирование текстов [3] и пр. Имеются два основных подхода к её решению: обучение с учителем, когда алгоритм «обучается» задаваемым «учителем» категориям, и самообучение, когда алгоритм сам определяет, какие категории релевантны данному тексту. Первоначально речь шла о том, чтобы каждому тексту приписывалась единственная категория. В последнее время всё чаще допускается многоаспектная категоризация, когда один и тот же документ может сопровождаться многими категориями (multi-label classification). В частности, нас интересует проблема рубрикации документов, таких как научные публикации, в системе рубрик, заданных таксономией соответствующей области знания или технологии. Например, публикации в сфере информатики и вычислительных процессов могут индексироваться рубриками так называемой Computing Classification System [4] – многоуровневой таксономии, разработанной международной Ассоциацией вычислительной техники (Association for Computing Machinery (ACM)) [5]. Мы будем обозначать эту таксономию через ACM CCS. Как и многие другие классификации, она представляет собой иерархическую систему, в которой каждая рубрика является частью более общей концепции и сама, в свою очередь, делится на более конкретные части. Например, согласно ACM CCS, «майнинг данных» – это часть «приложений информационных систем», в свою очередь содержащая такие части как «кластерный анализ» и «ассоциативные правила». В работе [6] приводятся обзор и результаты экспериментального сравнения методов многоаспектной категоризации с учителем для ситуаций, в которых категории образуют иерархическую систему, а в работе [7] подобный метод предлагается применительно непосредственно к системе рубрик классификации ACM CCS.

Тематика построения систем рубрикации в режиме самообучения практически не привлекала исследователей, вероятно, потому, что не существовало адекватного аппарата. Вообще, задачи анализа данных и текстов в режиме самообучения пока решаются с значительно более низкими уровнями точности, чем аналогичные задачи в режиме обучения с учителем (см., например, [8-10]). Данная работа посвящена исследованию возможности использования популярного в анализе текстов инструмента – мер релевантности строка-текст – для рубрикации документов в режиме самообучения. Использование мер релевантности строк и текстов в различных задачах обработки текстов насчитывает относительно долгую историю (см., например, [11-13]) и включает довольно развитый математический аппарат вероятностного моделирования применительно к проблематике извлечения информации ([11], [14]). Особенность данного подхода состоит в том, что используются только символьные последовательности и частоты их фрагментов, т.е. отсутствует какая-либо привязка к синтаксису, грамматике и семантике языка, на котором написаны тексты. С одной стороны, это определённое преимущество, так как методы, основанные на мерах релевантности, не зависят от особенностей языка и, следовательно, универсальны. С другой стороны, взятые как есть, они не могут учесть такие особенности естественного языка как наличие и использование синонимов, не говоря уже об особенностях структуры предложений.

Мы рассматриваем три основных подхода к измерению релевантности строка-текст, разработанные в международной литературе: (1) подход, основанный на векторном представлении текстов, идущий от самых ранних работ в области информационного поиска [15], [16]; (2) подход, основанный на вероятностной модели текстов и их тематики [12]; (3) подход, основанный на представлении текстов аннотированными суффиксными деревьями ([12], [17]). Мы дополняем этот последний подход оригинальной мерой релевантности, идея которой была сформулирована и описана нами в работе [13]. Эта мера отличается от других мер релевантности тем, что имеет чёткий операциональный смысл – суммарной условной вероятности символа в «совпадении», в сокращённой форме, СУВСС. Цель данной статьи – подвергнуть эти меры экспериментальному сравнению в проблеме рубрикации.

Работа структурирована следующим образом. В разделе 2 мы приводим определения рассматриваемых мер релевантности строка-текст. В разделе 3 рассматриваются наиболее популярные способы предобработки текстов. Раздел 4 посвящён описанию структуры проводимых экспериментов в разрезе трёх составляющих: (а) состав данных для обработки, (б) список используемых мер релевантности, (в) способы оценки результатов. Раздел 5 представляет результаты экспериментов. Раздел 6 заключает работу; в нём подытоживаются полученные результаты и формулируются направления дальнейшей работы.

Исследование осуществлено в рамках Научно-учебной группы «Методы визуализации и анализа текстов» Научного фонда НИУ ВШЭ (2011-2013 гг.). Авторы также выражают благодарность Академической программе за частичную поддержку работы, проделанной нами в рамках Международной научно-учебной лаборатории анализа и выбора решений и Научно-учебной лаборатории интеллектуальных систем и структурного анализа НИУ ВШЭ. Мы благодарны рецензенту за замечания, учтённые в процессе доработки статьи.

2. Меры релевантности

Мы рассматриваем три основных способа представления текстов: векторную модель, вероятностную модель и аннотированное суффиксное дерево АСД.

2.1. Векторная модель

Согласно векторной модели [15], текстовый документ представляется множеством слов (или каких-нибудь других элементов документа), причём каждому слову соответствует своя координата векторного пространства. В качестве значения обычно используется величина так называемой *tf-idf* кодировки, равная количеству вхождений слова в документ, делённому на логарифм относительного количества документов, содержащих это слово [15].

$$\text{Пусть } w_{ia} = tfi \cdot df = tf_{ia} \cdot log \frac{|A|}{n(t_i)+1},$$

где tf_{ia} – частота терма f_i в аннотации a, $n(t_i)$ – число аннотаций, содержащих терм f_i, $|A|$– общее число аннотаций. Пусть w_{ia}, w_{iq} – веса терма f_i в аннотации (*abstract*) $a \in A$ и таксономической теме (*topic*) q. Сходство между таксономической темой и аннотацией определяется по формуле:

$$relevance(topic, abstract) = cos(\vec{q}, \vec{a}) = \frac{\vec{q} \cdot \vec{a}}{\|\vec{q}\| \cdot \|\vec{a}\|} = \frac{\sum_{i=1}^{N} w_{ia} \cdot w_{iq}}{\sqrt{\sum_{i=1}^{N} w_{ia}} \sqrt{\sum_{i=1}^{N} w_{iq}}}.$$

2.2. Вероятностная модель

Вероятностная мера релевантности используется, в основном, в задачах извлечения /поиска информации. Она построена в предположениях теоретической модели, согласно которой каждый текстовый документ представляется как смесь двух Пуассоновских распределений [14]. Одно из них отвечает за распределение обычных слов, другое – за распределение «элитных» слов, то есть, тех, на которых лежит основная смысловая нагрузка в разрезе рассматриваемой тематики. Ставшая очень популярной в последнее время мера релевантности BM25 придаёт больший вес «значимым» термам и меньший – «незначимым»:

$$relevance(topic, abstract) = \sum_{i=1}^{N} IDF(t_i) \frac{(k_1+1)\, tf_{ia}}{tf_{ia} + k_1\left(1 - b + b\frac{|A|}{avgdl}\right)},$$

где *avgdl* – среднее количество слов в аннотации, a, b, k_1 – константы, равные, как правило 1.5 и 0.75, соответственно, согласно [14].

В качестве нормализующего сомножителя используется функция , имеющая смысл обратной частоты:

$$IDF(t_i) = log \frac{|A| - n(t_i) + 0.5}{n(t_i) + 0.5},$$

где $|A|$ – общее число аннотаций, а $n(t_i)$ – число аннотаций, содержащих терм t_i.

2.3. Модель аннотированного суффиксного дерева (АСД)

Согласно модели АСД [12, 13], текстовый документ характеризуется не совокупностью слов или термов, а фрагментами – последовательностями символов в том порядке, в котором они встречаются в тексте.

Аннотированное суффиксное дерево – это структура данных, используемая для вычисления и хранения всех фрагментов текста совместно с их частотами. Она задаётся как корневое дерево, в котором каждый узел соответствует одному

символу и помечен частотой того фрагмента текста, который кодирует путь от корня до данного узла.

Чтобы ограничить глубину конструируемого АСД, мы разбиваем текст на короткие фрагменты – «строки», состоящие из двух – четырех слов. Алгоритм построения АСД, представляющий собой модификацию известных методов построения суффиксных деревьев [12], [17], описан нами в работе [13] (*рис. 1*).

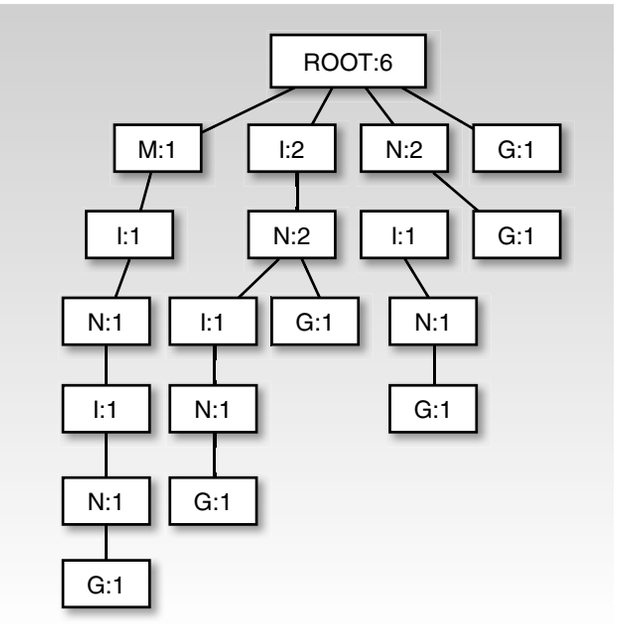

Рис. 1. Аннотированное суффиксное дерево (АСД) для строки «mining»

Оценка степени релевантности, или «присутствия» таксономической темы в данном АСД, вычисляется следующим образом:

1. Выделяются все суффиксы, т.е. конечные фрагменты, строки таксономической темы.

2. Для каждого суффикса вычисляется оценка его совпадения (match) с АСД:

$$score\,(match(suffix, ast)) = \\ = \sum\nolimits_{node \in match} \varphi\left(\frac{f(node)}{f(parent(node))}\right),$$

где совпадение – это путь от корня дерева, кодирующий совпадающий с ним префикс суффикса или суффикс целиком, $f(node)$ – частота, приписанная узлу АСД из совпадения, $f(parent(node))$ – частота, приписанная родителю данного узла.

3. Оценка релевантности вычисляется как среднее всех оценок, приходящееся на один символ:

$$relevance(topic, abstract) = SCORE(topic, ast) = \\ = \frac{\sum\nolimits_{suffix} score(match(suffix, ast)) \Big/ |\,suffix\,|}{|\,string\,|}$$

где $|suffix|$, $|string|$ – количество символов в суффиксе и в строке.

В этой формуле *score* – это одна из трех шкалирующих функций $\varphi(x)$, рекомендованных в [12]:

- ✦ линейная (*linear*), $\varphi(x) = x$
- ✦ логистическая (*logit*),

$$\varphi(x) = log\,log\frac{x}{1-x} = log(x) - log(1-x)$$

- ✦ квадратный корень (*root*), $\varphi(x) = \sqrt{x}$.

Из этих трёх только линейная, ничего не меняющая функция, имеет очевидный операциональный смысл – средней условной вероятности символа в совпадении (СУВСС); две нелинейные шкалы из [12] использованы для контроля.

Очевидно, что короткие элементы текста не могут нести особой тематической направленности. Поэтому возникает гипотеза, что вклады узлов начальных уровней АСД в оценки релевантности носят характер шума, и оценка релевантности станет более адекватной, если ее очистить от вклада узлов начальных уровней. Для проверки этой гипотезы мы обнуляли частоты узлов на первом, втором и т.д. уровнях от корня и обозначали такие способы вычисления через $\varphi.X$, где φ – вид шкалирующей функции, а X – уровень в АСД, до которого обнулялись частоты.

3. Способы представления текста

Использование векторной и вероятностных моделей предполагает представление текста в виде неупорядоченного набора термов. Под термами понимаются либо слова в исходном виде, либо некоторые значимые фрагменты слов, как правило, основы, часто называемые «стемами», либо же словарные формы (леммы) слов [2]. Кроме этих двух традиционных способов представления текста, мы рассматривали способ, согласно которому в качестве термов используются совпадения, получаемые при наложении всех ключевых словосочетаний на АСД. Конкретные способы выбора термов из множества всех слов или всех совпадений, использованные в экспериментах, представлены в *табл. 1.*

Таблица 1.

Способы представления текста как совокупностей термов

Обозначение	Описание
words	Все вхождения слов в неизменённом виде.
stems	Стемы (основы) всех слов. Для выделения стемов использован стеммер Портера [18] из библиотеки NLTK [19].
coll3	Все совпадения, полученные наложением всех таксономических тем на АСД для текстов, в качестве строк которых взяты последовательные тройки слов.
coll3.4	Те из термов coll3, которые состоят из 4 и более букв.
coll3.5	Те из термов coll3, которые состоят из 5 и более букв.
coll3.6	Те из термов coll3, которые состоят из 6 и более букв.
coll3_long	Те термы из coll3, которые являются самыми длинными из совпадений соответствующей таксономической темы с АСД.
coll3_long.4	Те из термов coll3_long, которые состоят из 4 и более букв.
coll3_long.5	Те из термов coll3_long, которые состоят из 5 и более букв.
coll3_long.6	Те из термов coll3_long, которые состоят из 6 и более букв.

4. Постановка эксперимента

Определим три основные составляющие вычислительного эксперимента:

(1) набор данных, на которых производится сравнение;

(2) набор методов, участвующих в сравнении;

(3) способ оценки качества результатов.

4.1. Выбор данных

Эксперимент проводился для коллекции данных, состоящей из трех частей: аннотаций научных статей, таксономии ACM CCS 2012, а также приписанных статьям их авторами тем из этой таксономии (*см. рис. 2*). Эти части кратко представлены ниже.

1. Аннотации всех научных статей, опубликованных за период с начала 2007 года по первый квартал 2013 года включительно, в следующих журналах, размещённых на портале ACM [5]:

a. ACM Transactions on Knowledge Discovery from Data (TKDD)

b. ACM Transactions on Internet Technology (TOIT)

c. ACM Transactions on Speech and Language Processing (TSLP).

Выбор журналов определялся профессиональными интересами авторов. Общее число аннотаций в данной коллекции – 244.

2. Таксономия ACM CCS 2012, состоящая из 2074 таксономических тем [4]. В таксономии ACM CCS 2012 6 уровней. На первом уровне располагается 13 основных разделов (*см. рис. 2*), на втором уровне – 90, на третьем – 547, на четвертом уровне находится большая часть листьев таксономии – 1074 тем.

3. Авторские темы, приписанные аннотации – это, как правило, 2-3 таксономические темы низших уровней таксономии, а также все темы, лежащие на пути от корня до них в дереве таксономии ACM CCS.

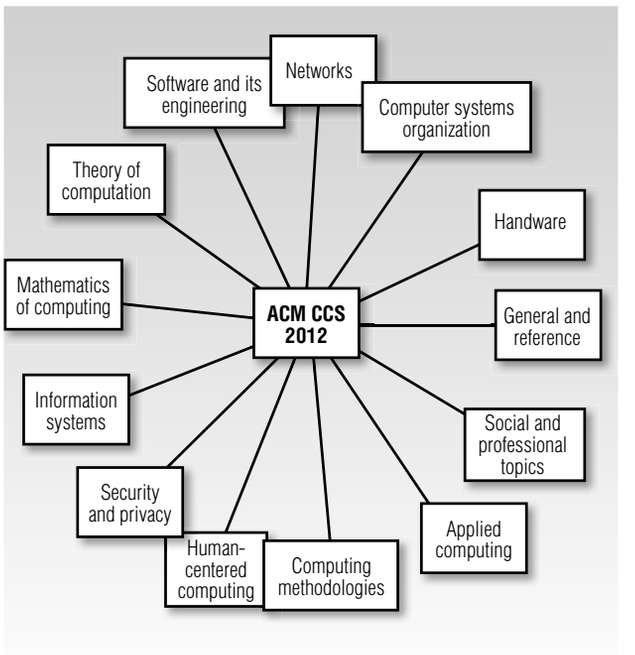

Рис. 2. Первый уровень таксономии ACM CCS 2012 [4]

Пример документа из рассматриваемого множества приведён в *табл. 2.*

Отметим, что авторы этой в своей рубрикации предпочли оттенить взаимодействие человека и компьютера, тогда как согласно аннотации, статья представляет собой скорее упражнение в применении вероятностной модели кластер-анализа для выявления сообществ. Термины «cluster» и «clustering» 6 раз участвуют в различных подразделениях таксономии ACM CCS, но никак не отражены в авторской рубрикации. Подобные нюансы интерпретации должны учитываться при оценке систем автоматической рубрикации.

Таблица 2.

Пример аннотации, участвующей в эксперименте. Статья выбрана случайно

Discovering Knowledge-Sharing Communities in Question-Answering Forums	
Mohamed Bouguessa, Shengrui Wang, Benoit Dumoulin	
ACM Transactions on Knowledge Discovery from Data (TKDD), V. 5, no.1, December 2010	
In this article, we define a knowledge-sharing community in a question-answering forum as a set of askers and authoritative users such that, within each community, askers exhibit more homogeneous behavior in terms of their interactions with authoritative users than elsewhere. A procedure for discovering members of such a community is devised. As a case study, we focus on Yahoo!Answers, a large and diverse online question-answering service. Our contribution is twofold. First, we propose a method for automatic identification of authoritative actors in Yahoo!Answers. To this end, we estimate and then model the authority scores of participants as a mixture of gamma distributions. The number of components in the mixture is determined using the Bayesian Information Criterion (BIC), while the parameters of each component are estimated using the Expectation-Maximization (EM) algorithm. This method allows us to automatically discriminate between authoritative and nonauthoritative users. Second, we represent the forum environment as a type of transactional data such that each transaction summarizes the interaction of an asker with a specific set of authoritative users. Then, to group askers on the basis of their interactions with authoritative users, we propose a parameter-free transaction data clustering algorithm which is based on a novel criterion function. The identified clusters correspond to the communities that we aim to discover. To evaluate the suitability of our clustering algorithm, we conduct a series of experiments on both synthetic data and public real-life data. Finally, we put our approach to work using data from Yahoo!Answers which represent users activities over one full year.	
Таксономические темы ACM CCS, приписанные автором (авторские темы)	
Human-centered computing	Information systems
Human computer interaction (HCI)	Information systems applications
Interaction paradigms	Data mining
Web-based interaction	

4.2. Выбор мер релевантности

В качестве мер оценки релевантности таксономической темы и аннотации научной статьи берутся популярные меры (*см. табл. 3*).

Таблица 3.

Обозначения рассматриваемых мер релевантности

Обозначение	Мера релевантности
cosine	косинусная мера релевантности
okapibm25	мера релевантности BM25
ast.linear	мера СУВСС с линейной шкалирующей функцией
ast.logit	мера СУВСС с логистической шкалирующей функцией
ast.root	мера СУВСС со шкалирующей функцией в виде квадратного корня

4.3. Оценка качества результатов

Мы использовали для оценки результатов две популярные характеристики точности: MAP (Mean Average Precision) и nDCG (normalized discounted cumulative gain) [11]. Они часто используются при разработке рекомендательных систем [20], систем извлечения новостей [21], обучении ранжированию [22, 23]. Для их вычисления может использоваться следующая общая схема отбора таксономических тем:

1. Таксономические темы ранжируются по убыванию их релевантности;
2. Отбираются первые k (топ **k**) таксономические темы, отсекая все остальные;
3. Вычисляется оценка получившегося ранжирования.

Мера MAP может быть представлена следующим образом:

$$AveP = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{|relevant\_topics|},$$

$$MAP = \frac{\sum_{a \in abstracts} AveP(a)}{|abstracts|},$$

где $P(k)$ – точность на уровне k в упорядоченном по убыванию меры релевантности списке таксономических тем, $rel(k)$ – бинарный показатель, принимающий значение 1, если k-тая таксономическая тема в списке является авторской, и 0 в обратном случае, $|relevant\_topics|$– число авторских таксономических тем, n – количество рассматриваемых таксономических единиц из топа списка. Здесь $AveP$ – средняя точность – рассчитывается для каждого текста рассматриваемого множества.

Мера $nDCG$ – это отношение оценки полученного ранжирования к оценке идеального случая:

$$nDCG_k = \frac{DCG_k}{IDCG_k}, \text{ где}$$

$$DCG_k = rel(1) + \sum_{i=2}^{k} \frac{rel(i)}{log_2 i} -$$

количество авторских таксономических тем среди топ k таксономических тем, нормированное на их место в ранжировании,

$$IDCG_k = rel(1) + \sum_{i=2}^{|relevant\_topics|} \frac{1}{log_2 i} -$$

значение «*DCG*» у идеального ранжирования. Мы выбрали $k = 15$ для *MAP* и *nDCG*, чего заведомо должно хватить, так как это k больше, чем количества авторских таксономических тем в нашей коллекции.

Кроме мер *MAP* и *nDCG*, использовалась также собственный критерий – количество публикаций, у которых авторские темы попали в топ k ранжированных таксономических тем. Будем обозначать эту меру через I_k (*Intersection at k*). Он позволяет легко

✦ отделить «хорошие» публикации – те, для которых удалось восстановить все или почти все авторские темы – от «трудных», для которых авторские темы находятся в конце соответствующего ранжирования, а также

✦ оптимальный порог отсечения k.

В принципе, меры *MAP* и *nDCG* тоже позволяют устанавливать пороговые значения, но они имеют значительно менее интуитивных характер, чем пороговые значения, которые определяются мерой I_k.

5. Результаты эксперимента

Полученные ранжирования тем оценивались по 6 характеристикам релевантности: четыре значения I_k, количество попаданий авторских тем в топ k, при k = 1, 5, 10, 15, а также меры *MAP* и *nDCG* при $k = 15$. Результаты оценки сведены в *таблицы* 4-6, соответствующие рассматриваемым мерам релевантности – косинусной (*табл. 4*), BM25 (*табл. 5*) и СУВСС (*табл. 6*).

Табл. 4 показывает, что косинусная мера в целом работает лучше всего на совпадениях из coll3 (полный список). Этот факт подтверждается как значениями *MAP*_15 и *nDCG*_15, так и значением меры I_k при всех рассматриваемых k. Однако различия с результатами на словах и основах (word и stem) не так уж и значительны.

Таблица 4.

Оценка полученных при использовании различных способов предобработки текстов результатов с помощью косинусной меры релевантности

Способ предобработки Слов	Количество попаданий авторских тем в топ k				MAP_15	nDCG_15
	I_1	I_5	I_10	I_15		
Words	10	44	60	73	0.0748	0.0245
Stems	8	37	57	77	0.0788	0.0250
coll3	14	41	58	76	0.0911	0.0278
coll3.4	14	31	46	59	0.0727	0.0207
coll3.5	9	31	50	71	0.0642	0.0237
coll3.6	12	31	45	57	0.0633	0.0218
coll3_long	8	32	48	60	0.0599	0.0187
coll3_long.4	7	22	41	56	0.0570	0.0182
coll3_long.5	7	12	51	65	0.0643	0.0208
coll3_long.6	8	32	46	53	0.0444	0.0158

Таблица 5.

Оценка полученных при использовании различных способов предобработки текстов результатов с помощью меры релевантности BM25

Способ предобработки Слов	Количество попаданий авторских тем в топ k				MAP_15	nDCG_15
	I_1	I_5	I_10	I_15		
Words	1	14	40	52	0.0631	0.0279
Stems	7	21	30	36	0.0869	0.0259
coll3	3	15	30	46	0.0524	0.0224
coll3.4	4	13	28	45	0.0532	0.0212
coll3.5	4	16	29	46	0.0577	0.0228
coll3.6	4	17	30	43	0.0547	0.0214
coll3_long	2	12	26	37	0.0446	0.0188
coll3_long.4	3	11	27	42	0.0482	0.0199
coll3_long.5	3	14	27	46	0.0540	0.0223
coll3_long.6	3	17	28	46	0.0528	0.0217

Табл. 5 выявляет двух победителей, одного при k = 1, 5, а другого – при k = 10, 15. В первом случае побеждает использование основ (stems), как по I_k, так и по *MAP*_15. Во втором случае слова (words) – наилучшие как по СУВСС, так и *nDCG*_15. Вместе с тем, нельзя не отметить, что все результаты использования меры BM25 хуже,

Таблица 6.

Оценка полученных результатов при использовании меры релевантности, основанной на АСД, при использовании различных видов шкалирующих функций

Вид шкалирующей функции и глубина очистки	Количество попаданий авторских тем в топ k				MAP_15	nDCG_15
	I_1	I_5	I_10	I_15		
linear.0	38	75	84	102	0.3588	0.1124
linear.1	35	77	89	105	0.3550	0.1133
linear.2	35	75	88	103	0.3486	0.1120
linear.3	34	70	90	105	0.3192	0.1020
linear.4	33	68	88	105	0.3059	0.0978
root.0	39	75	88	102	0.3657	0.1125
root.1	36	77	91	104	0.3561	0.1122
root.2	34	77	89	106	0.3497	0.1126
root.3	34	72	90	105	0.3232	0.1030
root.4	32	66	88	105	0.3064	0.0983
logit.0	7	36	48	57	0.1214	0.0450
logit.1	4	18	26	33	0.0521	0.0216
logit.2	6	29	46	56	0.0780	0.0335
logit.3	8	24	42	58	0.0850	0.0348
logit.4	8	37	54	64	0.1201	0.0482

чем соответствующие результаты использования косинуса. Особенно чётко это проявляется по оценкам I_k.

Табл. 6 показывает, что, как и ожидалось, нешкалированная, т.е. линейная, СУВСС является наилучшей по всем рассматриваемым критериям, причём наилучшие результаты достигаются на первом уровне «очистки».

Сравнение *таблиц* 4, 5, 6 показывает, что использование меры СУВСС, основанной на методе АСД, приводит к результатам, значительно превосходящим те, которые получаются при использовании других мер. Например, использование СУВСС порождает ранжирования, в которых на первое место попадает 35-39 авторских тем, тогда как использование косинуса и BM25 приводит максимум к 14 и 7 авторским темам на первых местах, соответственно. Значения традиционных критериев, MAP_15 и *nDCG*_15, для ранжирований по СУВСС превышают значения, достигнутые на ранжированиях по косинусной мере и BM25, в 4-5 раз.

При этом оказалось, что только однобуквенные сегменты оказались неинформативны.

Обратим внимание на то, что оценки качества рубрикации по СУВВС, хотя и намного лучшие, чем по другим мерам, но всё же не очень высоки. Например, они почти вдвое ниже, чем те, которые были достигнуты в уже упомянутой выше работе [6], где в общем и целом точность правильной рубрикации была порядка 70% текстов. Это можно объяснить двумя факторами. Во-первых, рубрикация в [6] делалась в режиме обучения с учителем, а не самообучения; общеизвестно, что результаты первого, как правило, лучше, чем второго. Во-вторых, при рубрикации статей с учителем используются не всевозможные, а только популярные рубрики, для которых процедуры классификации работают значительно точнее, чем для непопулярных. Если же обратиться к результатам, основанным на учете всех рубрик, то они вполне сопоставимы с нашими. В этом смысле показательны результаты победителей многочисленных соревнований по классификации текстовых материалов, приведённые в [25]. Например, при рубрикации категорий известной коллекции веб-страниц DMOZ (www.dmoz.org) участники различных соревнований (с фиксированными, и не очень большими, множествами категорий) показывали уровень успешности от 5% до 50% [25].

6. Заключение

Работа посвящена проблеме автоматизации рубрикации научных статей тематическими единицами таксономии соответствующей научной области. Так как данная проблема относится к области интерпретации ментальных, а не реальных объектов, то существенное значение приобретает адекватность эмпирического материала. Разметка множества публикаций тематическими единицами какой-либо иерархической классификации – непростое и не очень понятное дело. По нашему мнению, в качестве эмпирического материала лучше брать разметку, сделанную заинтересованными специалистами, чем заинтересованными дилетантами (см., например, использование DMOZ, иерархической системы вебсайтов, разработанной добровольцами, в [6]). Именно поэтому мы выбрали статьи, опубликованные в журналах, издаваемых наиболее авторитетной организацией в области информатики, ACM, и размеченные авторами согласно ACM CCS, классификации, разработанной именно этой

организацией. В этом плане мы в какой-то мере следовали работе [7], в которой тоже использовалась ACM CCS, хотя и в значительно более ранней версии 1998 г. Однако в работе [7] выбор публикаций и их рубрик оказался в какой-то мере случайным, так что подавляющая часть отобранных документов была помечена только одной рубрикой или вообще не помечена: менее чем 10% выбранных текстов оказались пригодными для рубрикации.

Мы использовали оценку релевантности строка-документ как основной механизм автоматизации рубрикации документов строками из заранее заданного списка в режиме самообучения. Было проведено сравнение трёх различных подходов к измерению релевантности: (а) косинусная мера векторной модели, (б) популярная мера вероятности порождения рубрик в рамках вероятностной модели, (в) средняя условная вероятность символа в совпадающих частях рубрики и текста на основе модели аннотированного суффиксного дерева. Оказалось, что в задаче рубрикации предложенная нами мера (в) с большим отрывом превосходит две другие, более популярные меры. Эффективность аппарата АСД отмечалась и в других приложениях, таких как категоризация [12]. Проверка гипотезы о том, что короткие, одно-, двух- и трех-буквенные сегменты текста не вносят полезного вклада в качество рубрикации, её подтвердила только в той части, которая относится к однобуквенным сегментам. Конечно, абсолютный уровень достигнутой точности остаётся относительно низким, что характерно и для других задач анализа текстов в режиме самообучения ([9, 10, 25, 26]).

Однако ситуация представляется не безнадёжной. Мы собираемся в будущем исследовать два пути дальнейшего развития. Первый – учёт синонимических отношений при оценке релевантности строка-текст. Второй путь связан с использованием латентного семантического анализа (LSA) [27] и/или аппарата латентных распределений Дирихле (ЛРД, LDA) [28] для вывода новых мер релевантности, основанных на многопараметрическом погружении пар строка-текст [29]. В настоящее время эти подходы используются только для рубрикации (multi-label classification) с помощью элементов самих анализируемых текстов; следует их адаптировать к задаче рубрикации текстов с помощью внешней системы рубрикации. ■

Литература

1. Сегалович И.В. Как работают поисковые системы // Мир Internet. 2002. №10.
2. Sebastiani F. Machine learning in automated text categorization // Journal of ACM Computing Surveys. 2002. Vol. 34, № 1, P.1–42.
3. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: МГУ, 2011.
4. ACM Computing Classification System 2012 // [Электронный ресурс]: http://www.acm.org/about /class/2012 (дата обращения 10.12.2013).
5. Association for Computing Machinery // [Электронный ресурс]: http://www.acm.org/ (дата обращения 10.12.2013).
6. Ceci M., Malerba D. Classifying web documents in a hierarchy of categories: a comprehensive study // Journal of Intelligent Information Systems. 2007. Vol. 28, № 1, P. 37–78.
7. Santos A.P., Rodrigues F. Multi-label hierarchical text classification using the ACM taxonomy // Proceedings of 14th Portuguese Conference on Artificial Intelligence (EPIA-2014). Aveiro, Portugal, October 12–15, 2010. P. 553–564.
8. Maetschke S., Madhamshettiwar P., Davis M., Ragan M. Supervised, semi-supervised and unsupervised inference of gene regulatory networks // Briefings in Bioinformatics. 2013. №5. P. 150–167.
9. Xu R., Morgan A., Das A. K., Garber A. Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon // Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. Stroudsburg, PA, USA, 2009. P. 63–70.
10. Grimmer J., Stewart B. M. Text as Data: The promise and pittfalls of automatic content analysis methods for political texts // Political Analysis. 2013. Vol. 21, № 3. P. 267–297.
11. Manning C.D., Raghavan P., Schütze H. An Introduction to Information Retrieval. –Cambridge: Cambridge University Press, 2008.
12. Pampapathi R., Mirkin B., Levene M. A suffix tree approach to anti-spam email filtering // Machine Learning. 2006. Vol. 65, № 1. P. 309–338.

13. Миркин Б.Г., Черняк Е.Л., Чугунова О.Н. Метод аннотированного суффиксного дерева для оценки степени вхождения строк в текстовые документы // Бизнес-информатика. 2012. № 3 (21). С. 31–41.
14. Robertson S., Zaragoza H. The probabilistic relevance framework: BM25 and beyond // Journal Foundations and Trends in Information Retrieval. 2009. Vol. 3, № 4, P. 333–389.
15. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing and Management. 1998. Vol. 25, № 5, P. 513–523.
16. Солтон Дж. Динамические библиотечно-поисковые системы: Пер. с. англ. М.: Мир, 1979. 557 с.
17. Gusfield D. Algorithms on strings, trees and sequences: computer science and computational biology. – Cambridge: Cambridge University Press, 1997.
18. Porter M.F. An algorithm for suffix stripping // Program: electronic library and information systems. 1980. Vol. 14, № 3. P. 130–137.
19. Bird S., Klein E., Loper E. Natural Language Processing with Python – Sebastopol: O'Reilly Media Inc, 2009.
20. Cantador I., Bellogin A., Vallet D. Content-based recommendation in social tagging systems // Proceedings of the fourth ACM conference on Recommender systems (RecSys-2010). Barcelona, Spain, September 26–30, 2010. P. 237–240.
21. Gupta A., Kumaraguru P. Credibility ranking of tweets during high impact events // Proceedings of the first Workshop on Privacy and Security in Online Social Media. Lyon, France, April 17, 2012. P. 2–8.
22. Xia F., Liu T., Wang J., Zhang W., Li H. Listwise approach to learning to rank - theory and algorithm // Proceedings of the 25th International Conference on Machine Learning (ICML-2008). Helsinki, Finland, July 5–9, 2008. P. 1192–1199.
23. Duh K., Kirchhoff K. Learning to rank with partially-labeled data // Proceedings of the 31st Annual International ACM Special Interest Group of Information Retrieval Conference (SIGIR-2008). Singapore, July 20–24, 2008. P. 251–258.
24. Valizadegan H., Jin R., Zhang R., Mao J. Learning to rank by optimizing NDCG measure // Advances in Neural Information Processing Systems. 2010. Vol. 22. P. 1883–1891.
25. Агеев М.С., Добров Б.В., Лукашевич Н.В. Автоматическая рубрикация текстов: методы и проблемы // Ученые записки Казанского государственного университета, серия Физико-математические науки. 2008. Т. 150, кн. 4. С. 25–40.
26. Galitsky B., Ilvovsky D., Kuznetsov S., Strok F. Matching sets of parse trees for answering multi-sentence questions // Proceedings of the Recent Advances in Natural Language Processing (RANLP-2013), Hissar, Bulgaria, September 12–14, 2013. P. 285–294.
27. Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. Indexing by Latent Semantic Analysis // Journal of American society for Infromation Science. 1980. Vol. 41, №6. P. 391–407.
28. Blei D. Probabilistic topic models // Communications of the ACM. 2012. Vol. 55, №4. P. 77–84.
29. Wang Q., Xu J., Li H., Craswell N. Regularized latent semantic indexing: A new approach to large-scale topic modeling // ACM Transactions on Information Systems. 2013. Vol. 31, № 1. P. 147–156.

USING PHRASE-TO-TEXT RELEVANCE SCORE TO ANNOTATE RESEARCH PUBLICATIONS

Ekaterina CHERNYAK,
Post-Graduate Student, Department of Data Analysis and Artificial Intelligence, School of Applied Mathematics and Information Science, Faculty of Business Informatics, National Research University Higher School of Economics

Address: 20, Myasnitskaya str., Moscow, 101000, Russian Federation
E-mail: echernyak@hse.ru

Boris MIRKIN,
Professor, Department of Data Analysis and Artificial Intelligence, School of Applied Mathematics and Information Science, Faculty of Business Informatics, National Research University Higher School of Economics

Address: 20, Myasnitskaya str., Moscow, 101000, Russian Federation
E-mail: bmirkin@hse.ru

Many semantic text analysis problems employ string-to-text relevance measures. Research paper annotation problem is no exception. In general, research papers are annotated according to a system of topics, organized as a taxonomy, a hierarchy of topics (or concepts). For example the papers, published in journals of the international Association of Computing Machinery (ACM), the most influential organization in the Computer Science world, are annotated according to the Computing Classification System taxonomy (ACM CCS).

String-to-text relevance measures should be used to automate the research paper annotation procedure since taxonomy topics are strings ant research papers or any of their constituents are texts. A relevance measure maps a string–text pair to a real number. The meaning of the mapping depends on the relevance model under consideration. Under any model, the higher the relevance value, the stronger the association between the string and the text.

This paper explores the use of phrase-to-text relevance measures to annotate research papers in Computer Science by key phrases taken from the ACM Computing Classification System. Three phrase-to-text relevance measures are experimentally compared in this setting. The measures are: (a) cosine relevance score between conventional vector space representations of the texts coded with tf-idf weighting; (b) a popular characteristic of the probability of «elite» term generation BM25; and (c) a characteristic of the symbol conditional probability averaged over matching fragments in suffix trees representing texts and phrases, CPAMF, introduced by the authors. Our experiment is conducted over a set of texts published in journals of the ACM and manually annotated by their authors using topics from the ACM CCS. Applying any of the relevance measures to an article results in a list of taxonomy topics sorted in the descending order of their relevance values. The results are evaluated by comparing these sorted lists and lists of topics assigned to articles manually. The higher a manually assigned topic is placed in a relevance based sorted list of topics, the more accurate the sorted list is. The accuracy of the computational annotations is scored by using three different scoring functions: a) MAP, b) nDCG, c) Intersection at k, where (a) and (b) are taken from the literature, and (c) is introduced by the authors. It appears, CPAMF outperforms both the cosine measure and BM25 by a wide margin over all three scoring functions.

Key words: phrase-to-text relevance, annotated suffix tree, text annotation, annotation scoring

References

1. Segalovich I.V. (2002) Kak rabotajut poiskovye sistemy [How search engines work]. Mir Internet, no 10. (in Russian)
2. Sebastiani F. (2002) Machine learning in automated text categorization. Journal of ACM Computing Surveys, vol. 34 no 1, pp. 1–42.
3. Lukashevich N.V. (2011) Tezaurusy v zadachah informacionnogo poiska [Thesauri for Infromation Retrieval]. Moscow, MGU. (in Russian)
4. ACM Computing Classification System 2012. Available at: http://www.acm.org/about /class/2012 (accessed 10.12.2013).
5. Association for Computing Machinery. Available at: http://www.acm.org/about /class/2012 (accessed 10.12.2013).
6. Ceci M., Malerba D. (2007) Classifying web documents in a hierarchy of categories: a comprehensive study. Journal of Intelligent Information Systems, vol. 28, no 1, pp. 37–78.
7. Santos A.P., Rodrigues F. (2010) Multi-Label Hierarchical Text Classification Using the ACM Taxonomy. Proceedings of 14th Portuguese Conference on Artificial Intelligence, (Aveiro, Portugal, October 12–15), pp. 553–564.
8. Maetschke S., Madhamshettiwar P., Davis M., Ragan M. (2013) Supervised, semi-supervised and unsupervised inference of gene regulatory

networks. Briefings in Bioinformatics, no 5, pp. 150–167.
9. Xu R., Morgan A., Das A.K., Garber A. (2009) Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (Stroudsburg, PA, USA), pp. 63–70.
10. Grimmer J., Stewart B. M. (2013) Text as Data: The Promise and Pittfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, vol. 21, no 3, pp. 267–297.
11. Manning C.D., Raghavan P., Schütze H. (2008) An Introduction to Information Retrieval, Cambridge: Cambridge University Press.
12. Pampapathi R., Mirkin B., Levene M. (2006) A suffix tree approach to anti-spam email filtering. Machine Learning, vol. 65, no 1, pp. 309–338.
13. Mirkin B.G., Chernyak E.L., Chugunova O.N. (2012) Metod annotirovannogo suffiksnogo dereva dlja ocenki stepeni vhozhdenija strok v tekstovye dokumenty [Annotated suffix tree method for estimating the string-to-text relevance]. Business Informatics, 2012, Vol. 3, no 3 (21), pp. 31–41.
14. Robertson S., Zaragoza H. (2009) The Probabilistic Relevance Framework: BM25 and Beyond. Journal Foundations and Trends in Information Retrieval, vol. 3, no 4, pp. 333–389.
15. Salton G., Buckley C. (1998) Term-weighting approaches in automatic text retrieval. Information Processing and Management, vol. 25, no 5, pp. 513–523.
16. Salton G. (1975) Dynamic library and information processing. Englewood Cliffs, N.J.: Prentice-Hall, Inc.
17. Gusfield D. (1997) Algorithms on strings, trees and sequences: computer science and computational biology, Cambridge: Cambridge University Press.
18. Porter M.F. An algorithm for suffix stripping. Program: electronic library and information systems, vol. 14, no 3, pp. 130–137.
19. Bird S., Klein E., Loper E. (2009) Natural Language Processing with Python. Sebastopol: O'Reilly Media Inc.
20. Cantador I., Bellogin A., Vallet D. (2010) Content-based recommendation in social tagging systems. Proceedings of the fourth ACM conference on Recommender systems (Barcelona, Spain, September 26 –30), pp. 237–240.
21. Gupta A., Kumaraguru P. (2012) Credibility Ranking of Tweets during High Impact Events, Proceedings of the first Workshop on Privacy and Security in Online Social Media (Lyon, France, April 17), pp. 2–8.
22. Xia F., Liu T., Wang J., Zhang W., Li H. (2008) Listwise approach to learning to rank - theory and algorithm. Proceedings of the 25th International Conference on Machine Learning (Helsinki, Finland, July 5–9), pp. 1192–1199.
23. Duh K., Kirchhoff K. (2008) Learning to rank with partially-labeled data. Proceedings of the 31st Annual International ACM SIGIR Conference (Singapore, July 20–24), pp. 251–258.
24. Valizadegan H., Jin R., Zhang R., Mao J. (2010) Learning to Rank by Optimizing NDCG Measure. Advances in Neural Information Processing Systems, vol. 22, pp. 1883–1891.
25. Ageev M.S., Dobrov B. V., Lukashevich N.V. (2008) Avtomaticheskaja rubrikacija tekstov: metody i problemy [Automated text annotation: methods and problems]. Proceedings of Kazan University, Natural Science Series, 2008, vol. 150, no 4, pp. 25–40.
26. Galitsky B., Ilvovsky D., Kuznetsov S., Strok F. (2013) Matching sets of parse trees for answering multi-sentence questions. Proceedings of the Recent Advances in Natural Language Processing (Hissar, Bulgaria, September 12 – 14), pp. 285–294.
27. Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. (1990) Indexing by Latent Semantic Analysis. Journal of American society for Infromation Science, vol. 41, no 6, pp. 391–407.
28. Blei D. (2012) Probabilistic topic models. Communications of the ACM, vol. 55, no 4, pp. 77–84.
29. Wang Q., Xu J., Li H., Craswell N. (2013) Regularized latent semantic indexing: A new approach to large-scale topic modeling. ACM Transactions on Information Systems, vol. 31, no 1, pp. 147–156.