Анализ результатов тестирования алгоритма София при решении задачи классификации коллекции правовых документов

© Наталия Осипова

СПбГУ
osipovanata@mail.ru

Аннотация

Статья посвящена информационно-поисковому алгоритму София и анализу полученных с его помощью результатов при классификации коллекции правовых документов. В статье дается описание алгоритма, приводится структура и функциональные возможности информационно – поисковой системы (ИПС) София, рассматриваются методики тестирования алгоритма и сделанные на их основе выводы.

Введение

Алгоритм София разработан в рамках совместного проекта, выполняемого факультетом Прикладной Математики – Процессов Управления Санкт Петербургского Государственного Университета и Лабораторией инженерии знаний (NIKEL) университета Ольстера (University of Ulster, UK). Подробное описание алгоритма и результаты некоторых экспериментов приведены в статье [1] и в заявке на патент [2].

Статья посвящена анализу полученных с помощью алгоритма результатов при классификации русскоязычной коллекции документов по заранее заданным темам. В качестве исходных данных для тестирования были использованы коллекция правовых документов, категории документов и обучающая выборка для каждой категории. Для разбиения документов по категориям применялись 3 различные версии алгоритма.

Тестирование результатов работы алгоритма проводилось Российским семинаром по Оценке Методов Информационного Поиска (РОМИП).

В статье используются следующие термины и определения.

Документ – произвольный именованный текст (статья, книга, публикация, сайт).

Словарь коллекции – слова, встречающиеся в документах коллекции.

Контекст слова – набор слов, встречающихся в документах, которые содержат данное слово. Число слов в контексте определяет размер контекста.

Кластер – группа близких документов, причем понятие близости документов основывается на анализе содержащихся в документах словах.

Вероятностное распределение контекста или документа – набор вероятностей слов для данного контекста (документа) (набор отношений частот слов в контексте (документе) к размеру контекста (документа)).

Энтропия – мера неопределенности, характеризующая общеупотребительность слова. Энтропия вычисляется на основании анализа контекста слова по вероятностному распределению контекста.

1. Алгоритм София

Алгоритм София обеспечивает разбиение множества документов на узкие по смыслу кластеры. Он состоит из следующих шагов:

Шаг 1. Выявление узких контекстов

Для каждого слова из словаря коллекции на основе анализа вероятностного распределения контекста слова вычисляется энтропия, которая тем больше, чем более общеупотребительным является слово. Выбирая слова с наименьшей энтропией, мы получаем наименее общеупотребительные, “тяжелые” слова, которые и задают узкие контексты. При выявления слов с узкими контекстами можно рассматривать всю коллекция документов или некоторую ее часть. Также можно выбросить из рассмотрения некоторое подмножество слов общего словаря коллекции. Найденные узкие контексты далее рассматриваются как подтемы, представленные в коллекции документов. Обычно для специализированных коллекций в несколько сот тысяч документов формируется 1000 и более таких подтем.

Шаг 2. Кластеризация документов.

Для построения кластеров вычисляются расстояния между всеми документами и найденными ранее узкими контекстами, причем расстоянием является дивергенция Дженсона-Шеннена между вероятностными распределениями контекстов слов и документов. При проведении жесткой кластеризации для каждого документа определяется единственный ближайший к нему узкий контекст. Возможно проведение мягкой кластеризации, при которой документ направляется в несколько кластеров.

Построенная система кластеров используется для обеспечения семантического поиска по коллекции заданных документов. При этом в ответ могут входить документы, не содержащие в себе слов из запроса.

2. Структура и функциональные возможности ИПС София

На основе алгоритма София была реализована информационно поисковая система (ИПС) София.

Структурно ИПС состоит из следующих элементов:

- сервера базы данных
- WEB сервера
- сервера приложений
- клиентских рабочих мест

Документы и вся информация по кластеризации и поиску хранятся в базе данных (БД). Система управления БД (СУБД) обеспечивает хранение документов, реализацию алгоритмов кластеризации и поиска документов, доступ к БД пользователей системы.

В ИПС используется СУБД MS SQL Server 2000.

WEB сервер обеспечивает доступ пользователей к ИПС через Интернет или Интранет.

Клиентское рабочее место обеспечивает интерфейс для доступа пользователя к ИПС.

Сервер приложений обеспечивает работу клиентских рабочих мест. При реализации клиентских рабочих мест используются средства разработки C# и технология ASP.

БД ИПС содержит:

- документы
- словарь коллекции
- темы классификации
- контексты слов
- контексты тем

- кластеры

В ИПС решаются следующие задачи:

- кластеризация документов
- классификация документов
- прямой поиск
- контекстный поиск
- поиск близких документов для данного документа

Исходными данными для кластеризации являются документы и словарь документов. По исходным данным строится таблица контекстов слов. По таблице контекстов вычисляется энтропия каждого слова и отбираются слова с минимальной энтропией – узкоспециализированные термины, на основе которых и строятся кластеры.

Далее вычисляются расстояния между найденными узкоспециализированными терминами и всеми документами из коллекции. На основании расстояний происходит разбиение документов по кластерам

Для классификации коллекции документов с помощью обучающих выборок строятся контексты заданных тем. Множество документов разбивается на кластеры. Каждому кластеру ставится в соответствие одна или несколько тем.

3. Методика тестирования

Целями тестирования являлись:

- проверка корректности и эффективности работы алгоритма София
- сравнительная оценка различных методов, применяемых для кластеризации и классификации документов
- оценка скорости работы различных версий алгоритма София

Для тестирования алгоритма использовалась коллекция русскоязычных правовых документов. Объем коллекции составляет 65000 документов. При этом размер отдельных текстов в 85% случаев не превышает 1000 слов. В состав коллекции входили законодательные документы, относящиеся к различным тематикам.

Первый этап тестирования заключался в классификации документов по заданным темам. Для каждой категории существовал предопределенный набор релевантных для данной категории документов - обучающая выборка. Объем обучающей выборки составлял 5% от объема всей коллекции. Классификация проводилась с помощью трех версий алгоритма София. Было

применено два метода кластеризации и два метода классификации документов.

При разбиении коллекции документов на кластеры основной проблемой является большой объем информации, которую требуется обработать. При работе с коллекцией, для которой уже создан словарь терминов, достаточно не рассматривать слова, не вошедшие в словарь терминов. Если же словаря не существует (а так оно обычно и бывает), то необходимо выделить из всего множества слов коллекции некоторое рабочее подмножество и искать узкие контексты только среди этого подмножества.

При кластеризации применялось два метода уменьшения количества обрабатываемых слов. Первый метод (A1) заключался в том, что из рассмотрения были исключены слова, которые встречались более чем в 1000 документов, при объеме коллекции в 65000, т.е. в 1.5% всех документов. Такой прием обосновывается тем, что часто встречающиеся слова не могут быть узкоспециализированными, они усложняют дальнейшее вычисление энтропии слов и расстояний между узкими контекстами и документами. Границы рассматриваемых слов обычно определяются в зависимости от размера коллекции, количества слов и документов. Сужение словаря по этому методу позволило уменьшить время построения кластеров в 10 раз.

В основе второго метода (A2) кластеризации лежала идея построения контекстов слов без учета больших документов, которые значительно увеличивали время построения кластеров. При нахождении контекстов такие документы не рассматривались, хотя и не были удалены из коллекции. Так как таких документов было не более 15% от всего размера коллекции, то такой прием не мог существенно изменить результат формирования узких контекстов. Применение данного метода привело к уменьшению времени кластеризации в 6 раз.

Для классификации документов применялось два метода.

По первому методу (B1) кластер соотносился теме, если в нем содержались документы, которые принадлежали этой теме в соответствии с ее обучающей выборкой.

По второму методу (B2) вычислялось расстояние между кластерами и контекстами тем. Далее кластеру соотносилось 1-3 наиболее близких ему темы.

Таким образом для тестирования использовались три версии ("дорожки") алгоритма.

В первой (textan) и второй (docsan) версиях алгоритма кластеризация документов проводилась по A1.

Классификация первой версии (textan) проводилась по методу B1, а второй версии (docsan) – по методу B2.

В третьей версии (dicsan) кластеризация проводилась по методу A2, классификация по методу B2.

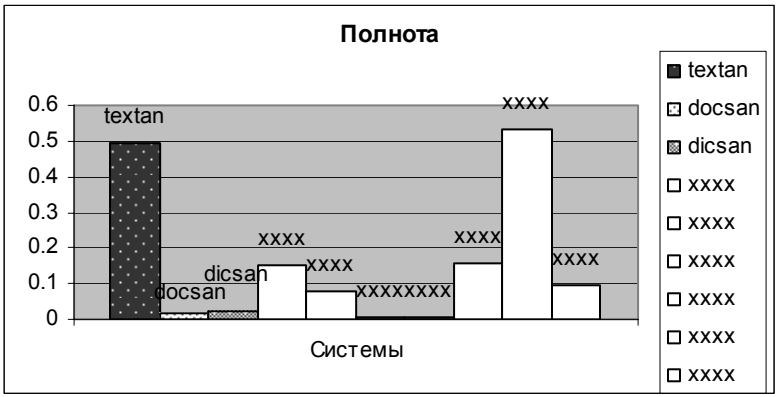
4. Анализ результатов тестирования

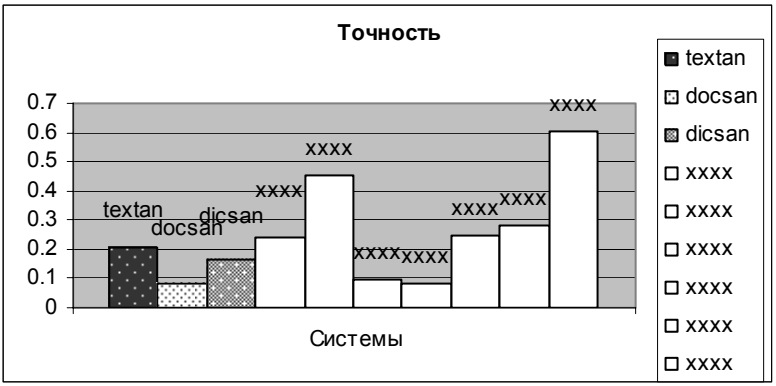
Тестирование результатов классификации документов проводилось независимыми экспертами по методике и с использованием программного обеспечения РОМИП.

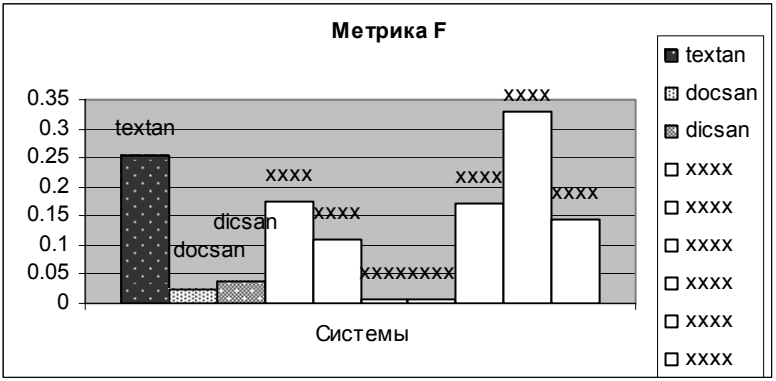
Из всего множества тем случайным образом выбиралось некоторое подмножество категорий. С помощью инструмента оценки РОМИП результаты, полученные экспертами, сравнивались с результатами, полученными различными ИПС. Были получены оценки (метрики) таких параметров системы как точность и полнота, и была построена мера F, объединяющая метрики полноты и точности.

Полнота показывает, какой процент верных документов было найдено системой. Точность – это отношение числа правильно найденных документов к общему числу найденных системой документов.

Были получены сравнительные результаты классификации коллекции как дорожек относительно друг друга, так и относительно других систем. Результаты приведены на графиках (1,2,3). Прочие ИПС, участвовавшие в классификации документов, обозначены как ‘xxxx’.







Оценка систем проводилась по следующим случайно выбранным категориям.

№	Категория
1	Семейное право
2	Наследственное право
3	Водное хозяйство
4	Общественное питание
5	Бытовое обслуживание населения
6	Арендные отношения
7	Международное космическое право
8	Территория в международном праве
9	Субъекты внешнеэкономических отношений

10	Внешнеэкономические сделки
11	Зоны свободной экономической торговли. Таможенные союзы

Результаты полноты для каждой темы приведены в таблице. Результаты София, которые оказались лучшими по данной категории, выделены жирным шрифтом. Все показатели приведены в процентах.

С \ К	1	2	3	4	5	6	7	8	9	10	11
textan	33	34	35	**60**	46	26	27	**98**	**75**	25	**100**
docsan	1	0	0.2	3	4	0	0.9	0	3	0	2
dicsan	0	0	4.3	2.3	0	5	0.9	8	3	0	0.8
xxxx	55	86	75	19	59	51	80	0	41	82	0
xxxx	21	39	2	22	15	6	0	1.4	0	5	0
xxxx	40	43	16	11	25	23	10	1.4	1.2	5	0
xxxx	23	4	2.5	1.1	18	7	0.9	0	1.2	10	0
xxxx	2.7	0	0	0	1.5	0	0	0	0	0	0
xxxx	2.2	0	0	0	1.5	0	0	0	0	0	0
xxxx	37	21	12	22	18	27	51	0	0	0	0

Процент верно найденных первым методом (textan) документов составляет около 50%, а у второго (docsan) и третьего (dicsan) методов – 1 - 4%. Так как методы кластеризации первого и второго методов были одинаковы, то отсюда следует, что на результат в значительной степени влияет используемый метод классификации. Для данной коллекции первый метод классификации оказался более эффективным.

Результаты показывают, что алгоритм София возвращает избыточное число документов, среди которых присутствуют почти все необходимые. Так как в системе София предусмотрена сортировка документов по релевантности заданной теме, то возможно представление результатов в виде, когда наиболее релевантные документы будут находиться в начале списка.

Заключение

Тестирование показало целесообразность дальнейшего развития математического и программного обеспечения ИПС София.

В данный момент ведется разработка усовершенствованного текстового обработчика на основе алгоритма Портера. Также разрабатываются новые алгоритмы поиска и классификации коллекций документов.

Литература

[1] Dobrynin V., Patterson D, Rooney N., Contextual Document Clustering, Lecture Notes in Computer Science № 2997, Advances in Information Retrieval, 2004 year, p.167-180.

[2] Dobrynin V., Patterson D, Rooney N. Galushka. N., UK Patent Application № 0322600.8.UK Patent Office 25.09.2003

The SOPHIA algorithm's analysis while solving the problem of legal documents' collection classification

Nataly Osipova
St.Petersburg State University

Annotation

This report is devoted to the information retrieval algorithm SOPHIA. We'll look through several methods which where applied to the legal documents' collection classification and try to analyze received results. In the report there are given the algorithm description, the structure and functionality of the SOPHIA system.