Разработка методов автоматического индексирования текстов на естественном языке для информационно-поисковых систем

© Алексей Николаевич Бевзов

Институт автоматики и электрометрии СО РАН
bvz@iae.nsk.su

Аннотация

Рассматривается разработка методов автоматического индексирования текстов на естественном языке (ЕЯ), которые могут использоваться при создании информационно-поисковых систем (ИПС). Основное внимание уделяется методологическому подходу, предназначенному для определения ключевых слов (терминов) в тексте на ЕЯ и определения веса этих терминов. Суть подхода состоит в том, что текст рассматривается не как простой набор слов ("bag of words"), а учитывается положение слов в предложении.

1 Введение

В настоящее время задача автоматической обработки и поиска смысловой информации в текстах на естественном языке является весьма актуальной. В этой области мы сталкиваемся с парадоксальной ситуацией [1]. Суть ее в том, что с одной стороны, в связи с развитием компьютерных технологий и средств связи доступ к информации упрощается. С другой стороны, возможность точного поиска необходимой информации уменьшается ввиду все возрастающего потока самой информации. Существующие на сегодняшний день методы обработки и поиска информации не вполне отвечают современным требованиям.

Среди наиболее известных методов, используемых для поиска информации, можно отметить: 1) библиотечные классификации (УДК, ББК, Классификация Дьюи, Классификация конгресса США) [2]; 2) поиск информации с помощью поисковых машин (ПМ) Интернет с указанием ключевых слов в поисковом запросе [3]; 3) использование лингвистических процессоров для автоматизированной обработки информации в текстах на ЕЯ [1], [4], [5]. Такие методы обработки и поиска информации наряду с положительными свойствами имеют ряд недостатков принципиального характера.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

Ограничения библиотечных классификаций состоят в том, что, они требуют обязательного знакомства как со стороны администратора ИПС, так и со стороны пользователя, что существенно затрудняет ввод новых документов в поисковые системы, а также их поиск. Недостаток поиска информации по ключевым словам с помощью ПМ Интернет состоит в том, что ПМ часто выдают огромное количество ссылок, обрабатывать которые приходится вручную. Принципиальное ограничение лингвистических процессоров обусловлено их привязкой к тому языку, на котором ИПС может обрабатывать и производить поиск смысловой информации, т.к. в основе логики работы таких процессоров лежит знание о семантических и синтаксических конструкциях конкретных ЕЯ

Данная работа описывает методику и результаты проведенных исследований, направленных на развитие и поиск альтернативных методов обработки и поиска информации в текстах на ЕЯ. В основе этих методов лежат идеи синергетического подхода к изучению сложных явлений [6]. В качестве исследуемых образцов текстов были использованы описания паттернов проектирования, приведенные в работе [7]. Данное исследование возникло как результат развития ИПС [8], первоначально предназначенной для более удобной работы с паттернами проектирования, которые используются при разработке программного обеспечения.

2 Текст как сложный синергетический объект

Суть предлагаемого подхода заключается в том, чтобы посмотреть на текст как на некоторый сложный синергетический объект, развернутый в пространстве и времени и получить различные временные ряды (ВР) из обычного текста на ЕЯ, а затем исследовать полученные данные с целью поиска в этих текстах наиболее значимой семантической информации.

Для исследования текста с помощью различных алгоритмов синергетики [9] была разработана специальная методика предварительной обработки текста с целью получения различных ВР. Ниже приведены основные этапы этой обработки.

1) Нормализация текста. На этом шаге производилась «оцифровка» текста. Сначала из теста выбрасывались предлоги и каждому слову текста присваивался уникальный идентификационный номер.

2) Выделение повторяющихся слов (ПС) выполнялось для того, чтобы отслеживать динамику эволюции текста на всем его протяжении.

3) Определение всех предложений в тексте. Критерием наличия предложения являлись такие знаки препинания, как ". ? ! ;".

4) Определение границ областей относительно позиции гармонического центра (ГЦ). Для каждого предложения определялись значения ГЦ, соответствующие пропорции золотого сечения. В этой части методики мы опирались на идеи, представленные в [10]. Позиция ГЦ равна 0.618 от начала предложения и измеряется в словах. Кроме того, выделялись границы трех областей относительно позиции ГЦ.

5) Расчет параметров ГЦ для каждого слова. На этом этапе для каждого слова проводился расчет параметров этого слова относительно позиции ГЦ предложения.

После выполнения указанных этапов предварительной обработки были предложены различные модели получения ВР, которые в дальнейшем исследовались синергетическими методами с целью выявления наиболее значимой смысловой информации.

3 Построение различных моделей текста

Для получения из текстов различных моделей ВР использовалась различная интерпретация ряда параметров, как, например, определение наблюдаемой переменной (НП), т.е. ключевого слова, способ получения значений НП (ЗНП) и т.д. Идентификация каждой модели задавалась с помощью: 1) определения того, как получаются ВР в результате компьютерной обработки текста; 2) определения параметра time (измеряется в словах) – момент проведения измерения ЗНП. 3) определения параметра deltaT (измеряется в словах) – длина временного интервала от одного времени измерения до другого. 4) определения параметра N – количество точек ВР и как они получаются. 5) определения того, что понимается под НП; 6) определения того, как считаются ЗНП;

В целом было рассмотрено несколько десятков моделей обработки текста с целью получения различных ВР. Ниже приводится описание некоторых моделей.

1) Идея получения ВР состоит в том, чтобы проанализировать динамику повторяемости ПС на протяжении всего текста через равные заданные «временные интервалы». Под «временными интервалами» подразумеваются фиксированные фрагменты текста, измеряемые в словах.

2) Идея получения ВР состоит в том, чтобы проанализировать динамику попадания выбранного ПС в разные области ГЦ на протяжении всего текста в каждом предложении.

3) Идея получения ВР состоит в том, что нас интересует динамика смещения позиции выбранного ПС относительно позиции ГЦ в предложении на протяжении всего текста. При этом, если ПС попадает в центр области ГЦ, то ЗНП равно некоторому максимальному значению, и это значение уменьшается до нуля по мере движения к левой или правой границе предложения.

4 Обработка ВР с целью получения некоторых идентификационных характеристик текста на ЕЯ

Обработка ВР для каждой из моделей велась аналогично тому, как это можно сделать при исследовании различных физических объектов. В работе использовался один из алгоритмов обработки ВР, описанный в [6], а именно – алгоритм расчета фрактальной размерности реконструированного аттрактора для идентификации интересующего нас объекта.

Предполагалось, что единственной информацией, которой мы обладаем относительно исследуемого объекта (в нашем случае – текста), является ВР (1), полученный согласно описанной выше методике обработки текста

$$X_0(t_0),\ldots..X_0(t_i)\ldots.X_0(t_N), \qquad (1)$$

где

$X_0(t_0)$ – экспериментально измеренное значение какой-либо переменной X_0 исследуемого объекта в момент времени t_0.

$X_0(t_i)$ – экспериментально измеренное значение этой же переменной X_0 в момент времени $t_i = t_0 + \Delta t \cdot i$.

Из этого ВР, получались *n*-мерные вектора $X1, X2, \ldots XN$ для реконструкции аттрактора. Значения компонент этих векторов получаются из элементов исходного ВР так, как это показано в таблице 1.

Таблица 1. Получение векторов реконструкции из исходного ВР

Вектор $X1$	Вектор XN
$X_0(t_0)$	$X_0(t_N)$
$X_0(t_0+\Delta t)$	$X_0(t_N+\Delta t)$
…..	….
$X_0(t_0+(n-1)\cdot\Delta t)$	$X_0(t_N+(n-1)\cdot\Delta t)$

Для этих векторов можно получить интегральную корреляционную функцию аттрактора по формуле:

$$C(r) = \frac{1}{N^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \theta(r - |X_i - X_j|), \qquad (2)$$

где

N – количество точек ВР (точнее, количество векторов реконструкции)

i, j – номера векторов реконструкции ($1 - N$)

$\theta(x)$ – функция Хевисайда: $\theta(x)$ = 1 если x > 0, $\theta(x)$ = 0 если x < 0

Xi – точка (вектор) фазового пространства с координатами ($X_0(t_i)$, … $X_0(t_i + (n-1)\cdot\Delta t)$)

$|X_i - X_j|$ – расстояние между точками Xi и Xj

r – некоторая заданная величина.

Если с учетом (2) построить график зависимости $\ln(C(r))$ от $\ln(r)$, то размерность аттрактора (d) исследуемой динамической системы будет определяться наклоном зависимости $\ln(C(r))$ от $\ln(r)$.

С целью поиска наличия аттрактора исследуемого объекта использовался описанный в [6] алгоритм.

1) Получить несколько графиков для корреляционной функции (2), исходя из данного ВР (1), рассматривая последовательно возрастающие значения размерности фазового пространства n.

2) Посмотреть, как изменяется при возрастании n наклон этих графиков d вблизи начала координат (2).

3) Если величина d в зависимости от n выходит на плато выше некоторого относительно небольшого n, то представленная данным ВР система должна иметь аттрактор. В этом случае размерность d представляет собой минимальное число переменных, необходимых для моделирования поведения, соответствующего данному аттрактору.

Кроме построения графиков корреляционной функции, для исходных ВР строились также графики *хаотичных ВР (ХВР)*. Хаотичные временные ряды получались из исходных текстов, в которых хаотичным образом были перемешаны слова.

Критерием возможности использовать для индексирования тот или иной набор повторяющихся слов считалась ситуация, когда графики наклона корреляционной функции для исходного ВР и аналогичные графики для ХВР будут отличаться.

После получения различных ВР для указанных в п. 3 моделей, был проведен численный эксперимент, в котором эти ВР были обработаны согласно описанной выше методике.

5 Результаты экспериментов

Ряд проведенных экспериментов. показал, что для некоторых моделей (модель 1) графики, полученные в результате обработки ВР и ХВР качественно не отличаются.

Для некоторых моделей (модификация модели 3) можно наблюдать качественное различие в поведении графиков обработки ВР и ХВР. Так, на рис. 1 и 2 представлены результаты обработки временных рядов, полученные для повторяющегося слова «адаптер» в тексте с описанием паттерна проектирования «Адаптер».

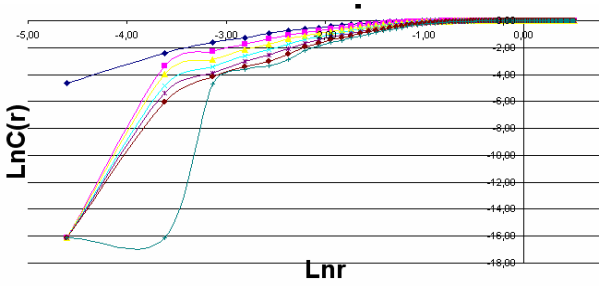

Рис. 1. Графики наклона корреляционной функции для исходного ВР.

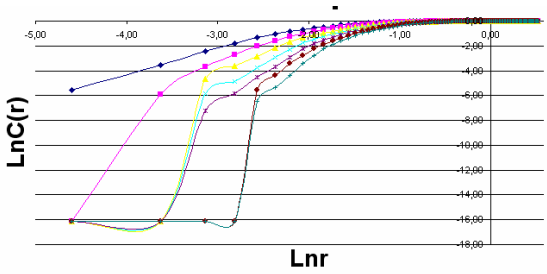

Рис. 2. Графики наклона корреляционной функции для хаотичного ВР.

Кроме указанного типа обработки ВР, опираясь на идею модели 3, был проведен расчет веса терминов (повторяющихся слов) документов, равных усредненной величине близости повторяющихся слов в тексте к положению ГЦ в предложении. После этого были проведены численные эксперименты для сравнения результатов поиска по предложенной методике расчета веса терминов с классическими методами расчета веса терминов.

Для выполнения процедуры поиска каждый документ был представлен в виде вектора в *N*-мерном евклидовом пространстве (*N* – количество терминов). Компоненты векторов соответствовали весу терминов, а коэффициент подобия между векторами документов и запросов считался по косинусной метрике.

Список терминов был образован из 70% наиболее частых слов с отсечением 15% наиболее частых и 15% наименее частых слов. Результаты проведенных экспериментов представлены в таблице 2. При этом коллекция текстовых документов состояла из описаний 23 паттернов проектирования, как они приведены в работе [7].

В качестве запросов использовались документы из коллекции документов в усеченном виде. Усеченный документ-запрос, которому необходимо было найти наиболее соответствующий ему документ в коллекции, формировался из 75, 50 и 25% исходного текста, начиная от начала. В таблице результаты поиска, соответствующие различным степеням усечения текста, разделены символом “/”.

Таблица 2. Результаты поиска документов

DP	tf (b)	tf	tf*idf	tf_ГЦ
1	15/21/21	1/2/4	1/8/7	1/12/11
2	6/6/6	1/2/9	1/1/4	1/1/2
3	10/10/14	1/1/1	1/1/2	1/1/5
4	3/13/13	1/2/2	1/1/1	1/1/1
5	5/5/5	1/1/1	1/1/1	1/1/1
6	12/8/18	1/1/1	1/3/3	1/2/3
7	4/4/5	1/1/1	1/1/1	1/1/13
8	11/10/10	1/1/6	2/1/2	1/1/3
9	3/3/3	1/1/2	1/1/2	1/1/1
10	9/8/13	1/1/1	1/1/2	1/1/1
Среднее	8.8/ 9.8/ 10.9	1.0/ 1.3/ 2.8	1.0/ 1.9/ 2.5	1.0/ 2.2/ 5.1

При этом в столбцах таблицы использованы следующие обозначения:
DP – порядковый номер паттерна в работе [7];
tf (b) – бинарный расчет веса термина (1 – слово присутствует в документе, 0 – отсутствует);
tf – расчет веса термина только исходя из его частоты употребления в документе;
tf*idf – классический расчет веса термина с учетом его частоты как в одном документе, так и во всей коллекции документов;
tf_ГЦ – усредненная величина близости повторяющихся слов к положению ГЦ в предложении (1 – если слово находится в положении ГЦ и линейное убывание до 0 на границах предложения).
В столбцах tf(b), tf, tf*idf, tf_ГЦ приведен ранжированный порядковый номер «правильного» документа, полученного в результате поиска, т.е. чем ниже этот номер, тем точнее выполнен поиск.

6 Заключение

Предложенная методика выявления ключевых слов в тексте и проведенные по этой методике численные эксперименты, позволяют говорить об обнадеживающих результатах данного подхода, поскольку для некоторых моделей текста поведение текста, содержащего смысл, отличается от поведения хаотически перемешанного текста. Проведенные эксперименты по поиску документов в коллекции документов позволяют говорить о сопоставимости результатов поиска в случае расчета веса терминов по предложенной методике и с использованием классических методов поиска.

Литература

[1] Краюшкин Д.В. Анализ технологий предварительной обработки документальной информации // Системы и средства информатики. – Вып. 15 / под ред. чл.-кор. РАН И.А. Соколова. – М., 2005.

[2] Мидоу Ч. Анализ информационно-поисковых систем.– М., 1970.

[3] Адамович, И.М. Поиск информации в WEB. Сравнительная оценка поисковых машин / И.М. Адамович, М.Ю. Заикин, Д.В. Земсков, А.Н. Пешков // Системы и средства информатики. – Вып. 13 / под ред. чл.-кор. РАН И.А. Соколова. – М., 2003.

[4] Козеренко Е.Б. Методы категоризации в построении многоязычного лингвистического процессора // Системы и средства информатики. – Вып. 15 / под ред. чл.-кор. РАН И.А. Соколова. – М., 2005.

[5] Кузнецов И.П. Особенности обработки текстов естественного языка на основе технологии баз знаний. // Системы и средства информатики. – Вып. 13 / отв. ред. чл.-кор. РАН И.А. Соколов. – М., 2003.

[6] Пригожин И., Стенгерс И. Познание сложного. – М., 2003.

[7] Гамма Э., Хелм Р., Джонсон Р., Влиссидес Дж. Приемы объектно-ориентированного проектирования. Паттерны проектирования. – СПб., 2001.

[8] Бевзов А.Н. Использование экспертных систем в задачах проектирования программного обеспечения // Датчики и системы. – 2002, № 12.

[9] Малинецкий Г.Г., Потапов А.Б. Современные проблемы нелинейной динамики. – М., 2002

[10] Москальчук Г.Г. Структура текста как синергетический процесс. – М., 2003.

Development of automatic indexation methods for natural text processing for data retrieval systems

Alexey Bevzov

The development of NLP (natural language processing) methods that can be used for automatic indexation in DRS (data retrieval systems) is considered. Main attention is given to methodological approach of extracting key words (terms) and their weight definition from natural text. The essence of the proposed approach is that text is considered not as a “bag of words”, but words position in clause is taken into account.