Интернет-журнал «Наукovedение» ISSN 2223-5167 <http://naukovedenie.ru/>

Том 7, №1 (2015) <http://naukovedenie.ru/index.php?p=vol7-1>

URL статьи: <http://naukovedenie.ru/PDF/70TVN115.pdf>

DOI: 10.15862/70TVN115 (<http://dx.doi.org/10.15862/70TVN115>)

УДК 004.85

Молчанов Артем Николаевич

ГКОУ ВПО «Академия Федеральной службы охраны Российской Федерации»

Россия, Орел

Сотрудник

E-mail: Tom_Silence@mail.ru

Скурнович Алексей Валентинович

ГКОУ ВПО «Академия Федеральной службы охраны Российской Федерации»

Россия, Орел

Доцент кафедры

Кандидат технических наук

E-mail: alexeymail2010@mail.ru

Математическая модель текста на естественном языке, учитывающая свойство когерентности

Аннотация. В настоящей статье приводится описание математической модели текста на естественном языке, размещенного на веб-странице информационного портала. Обосновывается актуальность разработки такой модели, а также выбор моделируемого свойства текста. Подробно раскрывается содержание этапов построения модели, приводятся практические результаты ее проверки на адекватность.

В частности, в статье приведены недостатки современных программных средств, осуществляющих сбор текстовой информации в сети Интернет. Показано, что задачу автоматического сбора информации с информационного портала можно свести к задаче классификации его веб-страниц на два взаимоисключающих класса: навигационного и информационного. Решение последней авторы связывают с моделированием свойства когерентности текста, размещенного на веб-странице. Продемонстрировано, что устанавливая с помощью предварительно обученного классификатора факт утраты смысловых связей между отдельными фрагментами текста, можно в автоматическом режиме эффективно отличать веб-страницы первого и второго классов.

На основе проведенных исследований автор делает вывод, что описанная в статье модель может служить теоретической основой для разработки альтернативного метода извлечения текстовой информации из сети Интернет. При этом отмечается, что такой метод, в отличие от известных, будет обладать устойчивостью к изменениям формы представления информации на веб-страницах информационного портала.

Ключевые слова: веб-страница; математическая модель текста; когерентность текста; распознавание текста; математическая лингвистика; количественный подход; параметры текста.

Ссылка для цитирования этой статьи:

Молчанов А.Н., Скурнович А.В. Математическая модель текста на естественном языке, учитывающая свойство когерентности // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 7, №1 (2015)
<http://naukovedenie.ru/PDF/70TVN115.pdf> (доступ свободный). Загл. с экрана. Яз. рус., англ. DOI:
10.15862/70TVN115

В последние годы такой феномен как «текст на естественном языке» все чаще становится объектом изучения не только лингвистики, но и некоторых естественных наук, таких как математика. Возросший интерес к изучению текста со стороны «нетрадиционных» для такого рода объекта наук, связывают с бурным развитием современных технологий информационного поиска текстовой информации, применяемых в различных информационных системах (ИС).

Первичной по отношению к частным задачам поиска информации является задача формирования контента ИС. При формировании контента системы встает вопрос об автоматизированном извлечении текстовой информацией из доступных для обработки источников. В подавляющем количестве случаев в качестве последних выступают информационные ресурсы сети Интернет (информационные порталы новостных агентств, аналитических и научно-исследовательских центров).

Одной из основных форм представления текстовой информации во всемирной сети является веб-страница. В свою очередь, информационный портал можно представить в виде совокупности взаимосвязанных веб-страниц двух типов: навигационных и информационных. Первый тип страниц предоставляет пользователю возможность быстрой и удобной навигации по структуре сайта, второй тип страниц – непосредственно саму информацию. Как показывает практика, такая структура сайта оптимальна для пользователей, но не оптимальна для автоматических средств поиска информации, поскольку ставит перед ними дополнительную задачу, связанную с распознаванием текста веб-страниц. Текст навигационных веб-страниц, в отличие от информационных, не содержит данных об описываемых событиях или явлениях в полном виде, а, следовательно, не может быть включен в контент ИС. Это означает, что модуль ИС, отвечающий за поиск страниц на сетевых ресурсах, должен обладать возможностью отличать такие страницы от информационных и заблаговременно удалять их из хранилища полученных данных.

На сегодняшний день для автоматического поиска текстовой информации с заданных ресурсов сети Интернет, применяют оффлайн-браузеры или специализированные программы. Оффлайн-браузеры, в виду ограниченности их функционала, при формировании контента ИС не используются, так как позволяют пользователю получить лишь полную копию всех страниц портала. Для решения данной задачи в современных системах поиска сетевой информации используется, так называемые, программы-посредники, работа которых основана на применении шаблонов разбора страниц. Программа-посредник – это программа, которая с определенной периодичностью заходит на заданный пользователем информационный ресурс, и, переходя по гиперссылкам, находящимся на его страницах, извлекает определенные фрагменты информации только с тех страниц портала, для которых в базе данных программы-посредника есть соответствующий шаблон разбора страницы.

Разработку шаблонов осуществляет администратор программы на этапе ее настройки. В общих чертах эту процедуру можно описать следующим образом. На подготовительном этапе администратор интернет-паука анализирует способ построения портала (на языке веб-разработчиков верстку сайта). Для этого изучаются образцы всех типов страниц портала (основной страницы, страниц разделов и всех его подразделов). По результатам анализа элементов логической и визуальной разметки (как правило, в качестве языка разметки используется язык HTML или XHTML) для каждого типа страниц на языке регулярных выражений [1] разрабатывается собственный шаблон разбора страницы. Данный шаблон представляет собой файл формата XML, загружаемый в настройки интернет-паука, который в соответствии с регулярными выражениями шаблона, извлекает со страницы требуемые фрагменты информации. Применительно к новостным сообщениям, к таким фрагментам относятся текст новости, его заголовок, автор, источник, дата и время публикации. Результаты извлечения информации также хранятся в файле формата XML, который в последующем, на

этапе анализа полученной информации, может быть преобразован к любому другому удобному для информационного работника формату (DOC, DOCX, TXT, RTF, PDF, FB2 и т.п.).

Очевидными достоинствами такого метода получения информации является точность и полнота извлекаемой из сети информации, относительная простота реализации процесса извлечения. Благодаря этим достоинствам, метод активно используется в большинстве современных систем поиска сетевой информации [2, 3]. Вместе с тем, опыт использования систем, реализующих данный метод, выявил ряд их существенных недостатков. При большом количестве источников информации первичная настройка системы, осуществляемая в ручном режиме, превращается в достаточно трудоемкую по временным ресурсам задачу. Разработчики сайтов пользуются неограниченной свободой в выборе средств и способов верстки страниц. При этом дизайн сайта, то есть оформление его страниц, пусть не значительно, но достаточно часто меняется. Изменение элементов оформления страниц сайта делает недоступным его контент для системы поиска информации. Для обеспечения стабильной работы системы ее администратору требуется вести мониторинг изменения дизайна сайтов и участвовать в постоянной перенастройке шаблонов разбора страниц.

Таким образом, разработка новых методов извлечения информации, требующих минимального участия человека и способных устойчиво работать при любых изменениях дизайна сайта продолжает оставаться важной и актуальной задачей.

В настоящей статье описана математическая модель текста на естественном языке (ЕЯ), учитывающая свойство когерентности. Данная модель впервые формализует одно из существенных свойств текста. По нашему мнению, она может стать теоретической основой для разработки альтернативного метода извлечения текстовой информации из сети Интернет, использующего процедуру распознавания типа веб-страницы по наличию или отсутствию в тексте сообщения смысловых связей между отдельными предложениями. Благодаря этой процедуре данный метод будет отличаться от известных большей степенью автоматизации самого процесса извлечения, а также обладать устойчивостью к изменениям, вносимым в дизайн сайта.

Свойство когерентности текста. Априорный словарь классов веб-страниц.

В теории моделирования под моделью понимают объект-заместитель объекта-оригинала, обеспечивающий изучение некоторых свойств оригинала [4]. В процессе моделирования за объект-оригинал был принят текст публицистического стиля, размещенный на веб-странице информационного портала, изучаемым свойством которого стало свойство когерентности. Выбор этого свойства был обусловлен результатами сравнительного анализа наличия основных свойств текстуальности у текстов, размещенных на информационной и навигационной веб-страницах. В число проанализированных свойств вошли членимость, автосемантизация отрезков текста, прагматичность, когезия и когерентность.

Как показали результаты сравнения, текст навигационной и текст информационной страниц можно разбить на отдельные составляющие (например, предложения), многие из которых сохраняют свой смысл и вне текста (например, цитаты, описания). Это, в свою очередь, свидетельствует о наличии у текстов таких свойств как членимость и автосемантизация [5]. При этом как первый, так и второй текст, выполняют определенную практическую функцию, т.е. они прагматичны [6, 7]: текст информационной страницы информирует пользователя о событиях или явлениях окружающего мира, а текст навигационной – побуждает его к определенным действиям (например, к переходу по ссылкам на другие страницы сайта или сетевые ресурсы).

Как в первом, так и во втором случаях, тексты веб-страниц обладают когезией [8]. Это означает, что между словами в рамках отдельно взятого предложения текста сохраняются

грамматические и синтаксические связи (например, осуществляется согласование слов по числу, времени, роду и падежу). Однако в отношении выполнимости свойства когерентности [9, 10, 11] тексты сильно разнятся (рисунок 1).

Свойство текста	Наличие свойства	
	Текст информационной веб-страницы	Текст навигационной веб-страницы
Членимость	+	+
Автосематизация отрезков текста	+	+
Прагматичность	+	+
Когезия (связность)	+	+
Когерентность (цельность)	+	-

Рисунок 1. Наличие основных свойств текста, размещенного на информационной и навигационной веб-страницах (разработано авторами)

Когерентность означает, что между отдельными предложениями текста должны быть сохранены смысловые связи. Если для текста информационной веб-страницы это условие выполняется, то в случае с текстом навигационной веб-страницы – нет. Между отдельными предложениями текста навигационной страницы наблюдается разрыв смысловых связей, что обусловлено особенностями его формирования (рисунок 2).

Рисунок 2. Процедура формирования текста навигационной веб-страницы из начальных фрагментов текста n-различных информационных веб-страниц (разработано авторами)

Как видно на схеме, изображенной на рисунке 2, текст навигационной веб-страницы T_Σ формируется на основе начальных фрагментов текста z_{1n} n -го количества информационных страниц T_n^{kt} .

Формально, это процедуру можно записать в виде:

$$T_\Sigma = \bigcup_{i=1}^n Z_{1i}; \quad (1)$$

Очевидно, что текст, составленный из фрагментов разных по содержанию и теме информационных сообщений, не станет цельным по смыслу, а значит, не будет обладать свойством когерентности. Таким образом, когерентность – это единственное свойство из числа проанализированных свойств текстуальности, по которому можно отличить текст информационной от текста навигационной веб-страниц. Это означает, что формализация данного свойства, позволит разработать систему распознавания, способную все множество веб-страниц портала эффективно классифицировать на два взаимоисключающих класса: класс навигационных и класс информационных сообщений.

Квантитативный подход в математической лингвистике

В ходе моделирования решалась задача выбора наиболее подходящей формы математической абстракции текста для описания свойства когерентности. Для решения этой задачи были проанализированы существующие методы представления текста на ЕЯ, общая классификация которых приведена на рисунке 3.

Рисунок 3. *Общая классификация методов представления текста на ЕЯ (разработано авторами)*

За основу для моделирования ранее неформализованного свойства текста был выбран количественный подход, разрабатываемый в рамках математической лингвистики. Данный подход предлагает описывать текст не множеством слов (термов), а частями речи, которыми эти слова выражены. Этим он, как минимум по двум причинам, выгодно отличается от методов, которые предполагают представлять текст в виде термов или производных от них элементов (например, *N*-грамм или шинглов).

Во-первых, размерность множества частей речи на порядки меньше размерности множества слов. Этот факт существенно сокращает размер априорного словаря параметров текста, что упрощает вычислительную процедуру отбора наиболее информативных из них.

Во-вторых, в условиях отсутствия априорной информации о теме текста, размещенного на веб-странице, сложно определиться с набором конкретных слов словаря, оценка статистических характеристик которых позволила бы отнести веб-страницу к классу информационных или навигационных сообщений. В тоже время относительно частей речи, это сделать можно, поскольку появление той или иной части речи практически не зависит от темы сообщения, а напрямую связано со структурой самого текста и особенностями его формирования. В данном случае акцент делается не на содержание текста, а на форму его представления. Известно, что порядка 80 процентов информации содержится в первом абзаце новостной статьи, так называемом « лиде ». На практике это означает, что автор статьи для написания лида будет стараться использовать только информативные с точки зрения повествования части речи, способные в максимально сжатом виде передать информацию о новостном событии.

Результаты анализа существующих моделей текста, разработанных в рамках количественного подхода [12], позволяют сделать вывод о том, что на современном этапе развития науки и техники с помощью таких моделей удастся описать несколько ключевых характеристик текста, таких как морфологический тип языка [13], стиль [14, 15, 16, 17] и жанр [18, 19] (рисунок 4).

Модели количественного подхода	Элементы текста, используемые для вычисления количественных параметров	Исследуемая характеристика (свойство) текста
Индексы Гринберга (1960)	морфы	Морфологический тип языка текста
Индексы Кауфмана (1961), Мистрика (1967), Шайкевича (1968), Кожинной (1972)	части речи	Стиль текста
Индексы Журавлева (1988), Ягуновой, Поспеловой (2014)	части речи	Жанр текста

Рисунок 4. Модели, разработанные в математической лингвистике в рамках количественного подхода (разработано авторами)

Обобщение положительного опыта использования частей речи для описания стилистических и жанровых характеристик текста, дало возможность адаптировать данный подход для описания свойства когерентности.

Математическая модель текста, учитывающая свойство когерентности

Аналитическая запись модели представлена группой выражений 2 – 6.

Текст веб-страницы можно представить в виде множества слов, мощность которого равна их количеству

$$T = \{t_1, t_2, \dots, t_i, \dots, t_w\}; \quad w = |T| \quad (2)$$

где T – множество слов текста на ЕЯ, t_i – i -ое слово текста, w – количество слов в тексте.

В свою очередь, все множество слов текста можно представить в виде объединения k -подмножеств различных частей речи, при этом каждое слово текста может быть отнесено к одному из этих подмножеств:

$$T = \bigcup_{j=1}^k C_j; \quad t_i \in C_j; \quad i = \overline{1, w}; \quad j = \overline{1, k}, \quad (3)$$

где C_j – подмножество слов j -ой части речи, k – количество частей речи в языке.

Для биективного отображения множества слов во множество частей речи X вводится функция $F(T)$. Результат работы этой функции есть вектор, i -ым элементом которого является количество i -ой части речи в описываемом тексте

$$F : T \rightarrow X, F(T) = (f_1(T), \dots, f_k(T)) = (x_1, \dots, x_k); \quad i = \overline{1, k}; i \in N, \quad (4)$$

где f_i – функция определения количества частей речи в тексте, x_i – количество i -ой части речи в тексте.

Вариант описания модели текста при утрате им свойства когерентности представлен выражениями (5) и (6).

Некогерентный текст можно записать в виде объединения n -разностей двух множеств: множества слов когерентного текста и множества слов текста, удаляемого из когерентного в результате формирования текста навигационной веб-страницы:

$$T_{\Sigma} = \bigcup_{i=1}^n (T_i^{\text{КТ}} \setminus T'_i); \quad (5)$$

где T_{Σ} – множество слов некогерентного текста (текста навигационной веб-страницы),

T^{KT} – множество слов когерентного текста (текста информационной веб-страницы),

T' – множество слов удаляемых из текста информационной веб-страницы при формировании текста навигационной веб-страницы,

n – количество информационных текстов, используемых для формирования текста навигационной веб-страницы.

Выражение (6) есть результат применения функции определения количества частей речи к выражению (5)

$$F(T_{\Sigma}) = F\left(\bigcup_{i=1}^n (T_i^{KT} \setminus T'_i)\right) = \sum_{i=1}^n (F(T_i^{KT} \setminus T'_i)) = \sum_{i=1}^n (F(T_i^{KT}) \ominus F(T'_i)) = \sum_{i=1}^n ((x_{i1}^{KT}, \dots, x_{ik}^{KT}) \ominus (\Delta x_{i1}, \dots, \Delta x_{ik})) \quad (6)$$

$$\ominus : \bar{x} \ominus \Delta x = (x_1, \dots, x_k) \ominus (\Delta x_1, \dots, \Delta x_k) = (x_1 - \Delta x_1, \dots, x_k - \Delta x_k)$$

где x_{ik}^{KT} – количество k -ой части речи в i -ом когерентном тексте, Δx_{ik} – количество k -ой части речи, удаляемое в i -ом когерентном тексте, $\ominus$ – операция поэлементного вычитания векторов.

Выражение (6) позволяет наглядно увидеть, что при утрате текстом свойства когерентности, численные значения его параметров будут меняться. Таким образом, сравнивая значения указанных параметров неизвестного текста с образцами параметров текстов, классовая принадлежность которых известна, можно оценить его когерентность.

Априорный словарь параметров текста для английского языка

На основе описанной модели был сформирован априорный словарь 27 параметров текста на английском языке. В его состав вошли параметры, доступные для измерения автоматическим морфоанализатором английского языка [20]. Эти параметры были сгруппированы в четыре группы: самостоятельные части речи, служебные части речи, формы частей речи, а также небольшая группа производных параметров (рисунок 5).

Самостоятельные части речи	Служебные части речи	Формы частей речи
существительное	предлоги	формы прилагательных
глагол	союзы	временные формы глаголов
наречие	артиклы	формы местоимений
прилагательное	инф. частицы	
местоимение	Производные параметры (соотношения отдельных частей речи, отношение уникальных слов к общему количеству слов в тексте)	
числительное		

Рисунок 5. Априорный словарь параметров текста (разработано авторами)

Данный словарь параметров обладает практически максимальным количеством информации об объекте распознавания в рамках выбранного подхода к представлению текста на английском языке. Однако для практического решения задачи распознавания, его размерность велика и должна быть обоснованно снижена. В этой связи стоит отметить, что сокращение числа параметров модели позволяет полностью или частично избежать серьезных проблем, связанных с вычислительной сложностью, а также с такими явлениями как «проклятие размерности» [21, 22] и эффект переобучения системы распознавания (также известного, как феномен Хьюджа [23]).

Известно, что при увеличении количества параметров модели объекта, сложность задачи его распознавания возрастает экспоненциально. Кроме того, в пространстве параметров с большей размерностью могут возникать явления, которые в пространстве с меньшей размерностью изначально не наблюдались. К числу таких явлений относится мультиколлинеранность [24], которая проявляется в форме вырожденности ковариационной матрицы параметров, что делает невозможным ее обращение в методах линейного и квадратичного дискриминационного анализа.

И, наконец, с ростом числа анализируемых параметров, классификатор в процессе обучения предельно точно настраивается на конкретную обучающую выборку, которая, в большинстве случаев, не является генеральной совокупностью объектов распознавания. Это, в свою очередь, означает, что попытки снизить величину ошибки классификатора до минимально возможных значений при распознавании объектов обучающей выборки могут привести к нежелательному росту числа ошибок распознавания на объектах тестовой выборки. В [25] доказывается, что при фиксированных значениях расстояния Махаланобиса между классами и объемах обучающей и тестовой выборок вероятность средней ошибки классификации растет с ростом количества анализируемых параметров объекта распознавания. В [23] можно найти такое описание этого явления: «в условиях фиксированного размера обучающей выборки, рост числа анализируемых параметров объекта распознавания, приводит к снижению точности его распознавания». В современных исследованиях при объяснении природы этого явления также делается особый акцент на ограниченность размера обучающей выборки. Так, в [26] утверждается, что переобучение классификатора носит фундаментальный характер и связано с неполнотой информации об объектах распознавания в момент принятия решения.

В качестве дополнительного аргумента, подтверждающего необходимость снижения размерности разрабатываемой модели и обобщающего приведенные выше доводы, можно привести принцип бритвы Оккама [27], который гласит, что «не следует множить сущее без необходимости». В машинном обучении данный принцип выражается в поиске наиболее простого описания объекта, позволяющего эффективно решать задачу его распознавания.

Процедура отбора информативных параметров из числа параметров априорного словаря

Известно, что из всех возможных вариантов построения модели объекта на основе априорного словаря параметров, предпочтительно выбрать ту модель, которой соответствует минимальное значение ошибки распознавания. Вместе с тем, данный подход не определяет порядок выбора параметров из всего множества доступных, что приводит исследователя к решению комбинаторной задачи, связанной с перебором всех возможных комбинаций признаков. Для того, чтобы избежать этой трудоемкой задачи, были проведены дополнительные исследования. Используя выборку из 3000 объектов, куда включались по 1500 представителей каждого класса, производился расчет расстояния в метрике Махаланобиса между классами по каждому параметру априорного словаря в отдельности. На рисунке 6 показаны значения расстояний Махаланобиса S между двумя распознаваемыми классами, вычисляемого на основе каждого анализируемого параметра текста априорного словаря.

Рисунок 6. Значения расстояния Махаланобиса по каждому параметру текста априорного словаря в отдельности (разработано авторами)

Данная диаграмма позволяет увидеть не только какой разделяющей способностью обладает каждый параметр текста в отдельности, но и, что важно, задать порядок составления комбинаций для оценки величины средней ошибки наивного байесовского классификатора $P_{\text{ош}}$. Исходя из представленных данных, целесообразно последовательно исключать из априорного словаря те признаки, которые обладают наименьшим расстоянием в выбранном метрическом пространстве, а, значит, являются менее информативными. Таким образом, при расчете значений средней ошибки $P_{\text{ош}}$ задачу полного перебора можно свести к задаче перебора только 27 комбинаций параметров.

Значения $P_{\text{ош}}$ для каждой из 27 комбинаций параметров при последовательном исключении менее информативных приведены на графике, представленном на рисунке 7.

Рисунок 7. Значения $P_{\text{ош}}$ Байесовского классификатора для 27 комбинаций параметров текста при последовательном исключении менее информативных (разработано авторами)

На первом шаге процедуры на вход классификатора Байеса подавались 3000 объектов обучающей выборки, описанные всеми 27 параметрами. Затем, после определения порога принятия решения по Байесовскому правилу, оценивалась средняя ошибка распознавания объектов тестовой выборки, включающая по 1000 объектов каждого класса в соответствии с выражением:

$$P_{ош} = \frac{N_{ош}^1 + N_{ош}^2}{N^1 + N^2}, \quad (7)$$

где $N_{ош}^1$, $N_{ош}^2$ – количество ошибочных решений при распознавании объектов первого и второго классов соответственно из тестовой выборки, $(N^1 + N^2)$ – общее количество объектов тестовой выборки. Таким образом, в результате использования при распознавании всех параметров текста значение $P_{ош}=0,16$ (рисунок 7, точка на графике, соответствующая по оси абсцисс значению "1-27").

На втором шаге при обучении классификатора и при распознавании объектов тестовой выборки был исключен наименее информативный параметр под номером 27, соответствующий параметру "Междометие" (рисунок 6). В результате, использовались признаки с 1 по 26 (рисунок 7). В данном случае $P_{ош}$ получилась равной 0,16 (точка на графике соответствует значению по оси абсцисс "1-26"). На третьем шаге при обучении классификатора и при распознавании использовались признаки с 1 по 25, т. е. был исключен наименее информативный параметр из оставшихся после второй итерации – "Артикль there" (рисунок 7). Таким образом, на каждом шаге последовательно исключались наименее информативные параметры.

В результате оценки $P_{ош}$ Байесовского классификатора на каждом шаге экспериментально было установлено, что при использовании в классификаторе для решения задачи распознавания большего количества признаков, чем 14 наблюдается эффект переобученности классификатора, когда добавление параметров не только не изменяет ошибку распознавания, но даже приводит к ее увеличению (рисунок 7, левая половина графика). В противоположность этому, если использовать небольшое количество наиболее информативных параметров, например первых 3, то наблюдается недообученность классификатора и как следствие большое значение ошибки при распознавании. Эксперименты показали, что минимальному значению $P_{ош}$ равной 0,11 соответствует комбинация признаков с 1 по 14. Это свидетельствует о том, что для эффективного распознавания веб-страниц достаточно включить в рабочий словарь признаков первые 14 (по расстоянию Махаланобиса) из 27 доступных для измерения параметров априорного словаря.

Установленный факт наглядно подтверждают отображения координат ста случайно выбранных навигационных и ста информационных веб-страниц в пространстве первых трех главных компонент, которые были получены в результате компонентного анализа исходного отображения объектов в 14-мерном пространстве параметров (рисунок 8).

Рисунок 8. Результаты сравнения объектов каждого класса в пространстве первых 3-х главных компонент (разработано авторами)

Выводы:

Результаты проверки разработанной модели свидетельствуют о ее адекватности: выбранные параметры чувствительны к изменению класса и могут быть использованы для решения задачи распознавания веб-страниц на основе количественной оценки свойства когерентности размещенного на них текста.

Дальнейшие исследования целесообразно проводить в плоскости разработки эффективного алгоритма распознавания, базирующегося на полученных теоретических и практических результатах моделирования.

ЛИТЕРАТУРА

1. Фридл Дж. Регулярные выражения, 3-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2008. – 608 с., ил.
2. Голубев С.А., Толчеев Ю.К. Шаров Ю.Л. Опыт внедрения и использования информационно-поисковой системы ODB-Text в Совете Федерации Федерального Собрания РФ// Современные технологии в управлении и образовании – новые возможности и перспективы использования. Сборник научных трудов. ФГУП НИИ «Восход», МИРЭА. – М., 2001. – с.58–61.
3. Подсистема сбора сообщений с сайтов новостей сети Интернет/ А.М. Андреев, Д.В. Березкин, В.В. Морозов и др.// Труды №1 молодых ученых, аспирантов и студентов «Информатика и системы управления». – М.: Изд. МГТУ им. Н.Э. Баумана. – 2003. – с.409 – 410.
4. Советов Б.Я., Яковлев С.А. Моделирование систем: Учебное пособие для вузов по специальности. "Автоматизированные системы обработки информации и управления". – 2-е изд., переработанное и дополненное. – М.: Высшая школа, 1998. – 319 с.
5. Головкина С.Х., Смольников С.Н. Лингвистический анализ текста – Вологда, 2006 г.
6. Гальперин И.Р. О понятии «текст». – Материалы научной конференции «Лингвистика текста», т. 1. - М., 1974. - С. 67.
7. Гальперин И. Р. Текст как объект лингвистического исследования. – М.: Наука, 1981.
8. Холлидей, М.А.К. Когезия в английском языке. М., 1976.
9. Москальская, О.И. Грамматика текста. – М.: Высшая школа, 1981.
10. Робер де Богранд и Вольфганг Дресслер Введение в лингвистику текста. – М., 1981.
11. Дейк ван Т.А., Кинч В. Стратегии понимания связного текста // Новое в зарубежной лингвистике. – Вып. 23: Когнитивные аспекты языка. – М., 1988.
12. Кашеева А.В. Квантитативные и качественные методы исследования в прикладной лингвистике // Социально-экономические явления и процессы. 2013. № 3 (049). С. 155-162.
13. Гринберг Дж. Квантитативный подход к морфологической типологии языков // Новое в лингвистике / Дж. Гринберг. 1963. Вып. 3. С. 60–94.
14. Кауфман С.И. Количественный анализ общезыковых категорий, определяющих качественные особенности стиля // Вопросы романо-германского языкознания, Коломна, 1961;
15. Мистрик Й. Математико-статистические методы в стилистике // Вопросы языкознания. – 1967. – № 3. – с. 42-52.
16. Шайкевич А.Я. Опыт статистического выделения функциональных стилей. – ВЯ. – 1968.– №1;
17. Кожина М.Н. О речевой системности научного стиля сравнительно с некоторыми другими. – Пермь, 1972;

18. Журавлев А.Ф. Опыт количественно-типологического исследования разновидностей устной речи // Разновидности городской устной речи. Сборник научных трудов. - М.: Наука, 1988. С. 84-150.
19. Поспелова А., Ягунова Е. Опыт применения стилистовых и жанровых характеристик для описания стилистовых особенностей коллекций текстов // Новые информационные технологии в автоматизированных системах: материалы семнадцатого научно-практического семинара. - М.: ИПМ им. М.В. Келдыша, 2014. - 560 с.
20. Интернет-ресурс <http://www.cll.khsu.ru/> – компьютерная лаборатория прикладной лингвистики Института информатики и телематики Хакасского государственного университета им. Н.Ф. Катанова.
21. Richard Ernest Bellman; Rand Corporation (1957). Dynamic programming. Princeton University Press. ISBN 978-0-691-07951-6., Republished: Richard Ernest Bellman (2003). Dynamic Programming. Courier Dover Publications. ISBN 978-0-486-42809-3.
22. Richard Ernest Bellman (1961). Adaptive control processes: a guided tour. Princeton University Press.
23. Hughes, G.F. (January 1968). "On the mean accuracy of statistical pattern recognizers". IEEE Transactions on Information Theory 14 (1): 55–63. doi:10.1109/TIT.1968.1054102.
24. Померанцев А. Классификация Российское хеометрическое общество, 2011 Электронный ресурс: <http://rcs.chemometrics.ru/Tutorials/classification.htm>
25. Грузман И.С., Киричук В.С., Косых В.П., Перетягин Г.И., Спектор А.А. Цифровая обработка изображений в информационных системах: Учебное пособие. – Новосибирск: Изд-во НГТУ, 2002. – 352 с.
26. Воронцов, К. В. Комбинаторная теория надёжности обучения по прецедентам: Дис. док. физ.-мат. наук: 05-13-17. – Вычислительный центр РАН, 2010. – 271 с. (<http://www.machinelearning.ru/wiki/images/b/b6/Voron10doct.pdf> , стр.63)
27. Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, Manfred K. Warmuth Occam's Razor Information Processing Letters 24(6): 377-380 (1987)
28. Schwarz, G. Estimating the dimension of a model. – Annals of Statistics. – 1978 T. 6. – 461--464 с.

Рецензент: Шишкин Николай Викторович, профессор кафедры, доктор технических наук, Академия ФСО России.

Molchanov Artem Nikolaevich

The Academy of the Federal Guard Service of the Russian Federation
Russia, Orel
E-mail: Tom_Silence@mail.ru

Skurnovich Aleksey Valentinovich

The Academy of the Federal Guard Service of the Russian Federation
Russia, Orel
E-mail: alexeymail2010@mail.ru

Mathematical model of natural language text for estimating coherence

Abstract. This paper describes mathematical model of natural language web-page text. It grounds actuality of model building and the text property decision. The paper writes up stages of model building and demonstrates model test results.

In particular the paper highlights drawbacks of up-to-date software tools for text data gathering in Internet. The authors of the paper offer alternative approach to the text data gathering procedure based on pattern classifiers. It is shown that the task of text data gathering could be turned into a web-page classification task. All the web-pages of the internet portal can be classified into two incompatible classes: navigational and informational. The solution of the classification task is associated with web-page text coherence modeling. It is shown that it is possible to classify web-pages automatically by means of trained classifier which can establish fact of text coherence loss.

With the results of survey the author draws a conclusion that the described model may be theoretical basis for developing of alternative method for the text data gathering in Internet. It is noted that in comparison with the known methods such method will be resistant to changing of web-page design.

Keywords: web-page; mathematical model of natural language text; coherence; text recognition; mathematical linguistics; quantitative approach; text parameters.

REFERENCES

1. Fridl Dzh. *Regulyarnye vyrazheniya*, 3-e izdanie. – Per. s angl. – SPb.: Simvol-Plyus, 2008. – 608 s., il.
2. Golubev S.A., Tolcheev Yu.K. Sharov Yu.L. Opyt vnedreniya i ispol'zovaniya informatsionno-poiskovoy sistemy ODB-Text v Sovete Federatsii Federal'nogo Sobraniya RF// *Sovremennye tekhnologii v upravlenii i obrazovanii – novye vozmozhnosti i perspektivy ispol'zovaniya*. Sbornik nauchnykh trudov. FGUP NII «Voskhod», MIREA. – M., 2001. – s.58–61.
3. Podсистема sbora soobshcheniy s saytov novostey seti Internet/ A.M. Andreev, D.V. Berezkin, V.V. Morozov i dr.// *Trudy №1 molodykh uchenykh, aspirantov i studentov «Informatika i sistemy upravleniya»*. – M.: Izd. MGTU im. N.E. Bauman. – 2003. – s.409 – 410.
4. Sovetov B.Ya., Yakovlev S.A. *Modelirovanie sistem: Uchebnoe posobie dlya vuzov po spetsial'nosti. "Avtomatizirovannye sistemy obrabotki informatsii i upravleniya"*. – 2-e izd., pererabotannoe i dopolnennoe. – M.: Vysshaya shkola, 1998. – 319 s.
5. Golovkina S.Kh., Smol'nikov S.N. *Lingvisticheskiy analiz teksta* – Vologda, 2006 g.
6. Gal'perin I.R. O ponyatii «tekst». – *Materialy nauchnoy konferentsii «Lingvistika teksta»*, t. 1. - M., 1974. - S. 67.
7. Gal'perin I. R. *Tekst kak ob"ekt lingvisticheskogo issledovaniya*. – M.: Nauka, 1981.
8. Khollidey, M.A.K. *Kogeziya v angliyskom yazyke*. M., 1976.
9. Moskal'skaya, O.I. *Grammatika teksta*. – M.: Vysshaya shkola, 1981.
10. Rober de Bogrand i Vol'fgang Dressler *Vvedenie v lingvistiku teksta*. – M., 1981.
11. Deyk van T.A., Kinch V. *Strategii ponimaniya svyaznogo teksta // Novoe v zarubezhnoy lingvistike*. – Vyp. 23: Kognitivnye aspekty yazyka. – M., 1988.
12. Kashcheeva A.V. *Kvantitativnye i kachestvennye metody issledovaniya v prikladnoy lingvistike // Sotsial'no-ekonomicheskie yavleniya i protsessy*. 2013. № 3 (049). S. 155-162.
13. Grinberg Dzh. *Kvantitativnyy podkhod k morfologicheskoy tipologii yazykov // Novoe v lingvistike / Dzh. Grinberg*. 1963. Vyp. 3. S. 60–94.
14. Kaufman S.I. *Kolichestvennyy analiz obshcheyazykovykh kategoriy, opredelyayushchikh kachestvennye osobennosti stilya // Voprosy romano-germanskogo yazykoznaniiya*, Kolomna, 1961;
15. Mistrik Y. *Matematiko-statisticheskie metody v stilistike // Voprosy yazykoznaniiya*. – 1967. – № 3. – s. 42-52.
16. Shaykevich A.Ya. *Opyt statisticheskogo vydeleniya funktsional'nykh stiley*. – VYa. – 1968. – №1;
17. Kozhina M.N. *O rechevoy sistemnosti nauchnogo stilya sravnitel'no s nekotorymi drugimi*. – Perm', 1972;
18. Zhuravlev A.F. *Opyt kvantitativno-tipologicheskogo issledovaniya raznovidnostey ustnoy rechi // Raznovidnosti gorodskoy ustnoy rechi*. Sbornik nauchnykh trudov. - M.: Nauka, 1988. S. 84-150.

19. Pospelova A., Yagunova E. Opyt primeneniya stilevykh i zhanrovyykh kharakteristik dlya opisaniya stilevykh osobennostey kollektsey tekstov // Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh: materialy semnadtsatogo nauchno-prakticheskogo seminara. – M.: IPM im. M.V. Keldysha, 2014. – 560 s.
20. Internet-resurs <http://www.cll.khsu.ru/> – komp'yuternaya laboratoriya prikladnoy lingvistiki Instituta informatiki i telematiki Khakasskogo gosudarstvennogo universiteta im. N.F. Katanova.
21. Richard Ernest Bellman; Rand Corporation (1957). Dynamic programming. Princeton University Press. ISBN 978-0-691-07951-6., Republished: Richard Ernest Bellman (2003). Dynamic Programming. Courier Dover Publications. ISBN 978-0-486-42809-3.
22. Richard Ernest Bellman (1961). Adaptive control processes: a guided tour. Princeton University Press.
23. Hughes, G.F. (January 1968). "On the mean accuracy of statistical pattern recognizers". IEEE Transactions on Information Theory 14 (1): 55–63. doi:10.1109/TIT.1968.1054102.
24. Pomerantsev A. Klassifikatsiya Rossiyskoe khemometricheskoe obshchestvo, 2011 Elektronnyy resurs: <http://rcs.chemometrics.ru/Tutorials/classification.htm>
25. Gruzman I.S., Kirichuk V.S., Kosykh V.P., Peretyagin G.I., Spektor A.A. Tsifrovaya obrabotka izobrazheniy v informatsionnykh sistemakh: Uchebnoe posobie. – Novosibirsk: Izd-vo NGTU, 2002. – 352 c.
26. Vorontsov, K. V. Kombinatornaya teoriya nadezhnosti obucheniya po pretседentam: Dis. dok. fiz.-mat. nauk: 05-13-17. – Vychislitel'nyy tsentr RAN, 2010. – 271 s. (<http://www.machinelearning.ru/wiki/images/b/b6/Voron10doct.pdf> , str.63)
27. Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, Manfred K. Warmuth Occam's Razor Information Processing Letters 24(6): 377-380 (1987)
28. Schwarz, G. Estimating the dimension of a model. – Annals of Statistics. – 1978 T. 6. – 461--464 s.