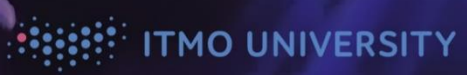
1-ая Международная научная конференция

Технологическая перспектива в рамках Евразийского пространства: новые рынки и точки экономического роста

20-21 ноября 2015 г.

Материалы конференции

г. Санкт-Петербург



ББК У9(2)80.3я43
УДК 339.94
Т38

Рецензенты:
Козловская Эра Анатольевна, доктор экономических наук, профессор кафедры инновационных и производственных систем Инженерно-экономического института Санкт-Петербургского политехнического университета Петра Великого
Яковлева Елена Анатольевна, доктор экономических наук, профессор кафедры экономики и финансов Санкт-Петербургского филиала Финансового Университета при Правительстве Российской Федерации

Т38 **Технологическая** перспектива в рамках Евразийского пространства: новые рынки и точки экономического роста. Материалы 1-ой Международной конференции / Под ред. проф. Н.С. Вороновой, В.С. Воронова, О.Н. Кораблевой, Ю.Е. Шелепина, А.М. Ельяшевича – СПб: Издательство НПК «РОСТ», 2015. - 245 стр.

ISBN 978-5-98217-087-3

Издание включает материалы 1-ой международной научной конференции «Технологическая перспектива в рамках Евразийского пространства: новые рынки и точки экономического роста»: развернутые тезисы и аннотации докладов.

ISBN 987-5-98217-087-3

Планируется провести анализ применения PLM систем в проектных, инжиниринговых и производственных компаниях как в России, так и за рубежом. В том числе, предполагается проведение исследовательской работы в направлении управления проектами в подобных компаниях, детальное изучение бизнес-процессов управления проектами компании и выявление путей повышения их эффективности.

Использование расширенного функционала планирования и шаблонов менеджера по управлению проектами, в условиях гибкости системы позволит компании автоматизировать деятельность по планированию ресурсов, и применить лучшие практики для основных проектов компании. Появится возможность вести проект совместно со всеми членами команды, что должно значительно повысить производительность участников проекта.

Организация управления проектами в компании, в частности, введение календарных планов работ, планов последовательности работ, технологий и методологий управления проектами, оптимизация загрузки ресурсов, коммуникаций по проекту и многое другое, даст наглядный пример использования возможностей системы по оптимизации работ над проектами компании. Организация работы бизнес-процессов управления проектами позволит увидеть преимущества использования системы во взаимодействии с другими программными средствами моделирования и оптимизации бизнес-процессов. Планируемый анализ и глубокое изучение вопроса, предоставит возможность, еще в процессе создания среды выявить множество проблем в работе компании, а на этапе разработанной среды предоставить их решение. Проделанная работа, в комплексе, позволит выявить пути повышения эффективности реализуемых проектов, за счет использования рассмотренных информационных технологий.

Литература:

1. Страница с описанием деятельности и основными достижениями Научно-Технического Центра «РОКАД» [Электронный ресурс]. – URL: http://rocad.ru (Дата обращения: 18.10.2015)
2. Описание PLM системы «ENOVIA» [Электронный ресурс]. – URL: http://www.3ds.com/ru/produkty-i-uslugi/enovia/ (Дата обращения: 22.10.2015)

Аналитический мониторинг. Метрики текстов на естественном языке.

Text Monitoring. The metrics of Natural Language Text.

С.В. Кулешов, Зайцева А.А.
Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации РАН,

Ключевые слова: онтология, аналитический мониторинг, обработка текстов, качество текстов

Key words: ontology, text monitoring, text processing, text quality

Введение

В работе предлагается сравнение метрик оценки свойств текстов на естественном языке применительно к подходу обработки текстов с применением автоматически формируемой в процессе работы онтологии. Подход является расширением методов лингвистической статистики и логико-статистических методов извлечения знаний. Основным отличием такого подхода от априорно-заданной онтологии, используемой в подходах, основанных на Semantic Web, является

автоматическое формирование семантического окружения (онтологии) в процессе работы системы [1].

Предлагаемый подход может использоваться в задачах семантического поиска документов, задачах мониторинга Интернет, а также для предоставления кратких аннотаций документов.

Для качественной ассоциативно-онтологической обработки текстов на естественном языке и адекватной работы систем аналитического мониторинга требуется решить подзадачи анализа свойств самих текстов и определения влияния их параметров на качество работы таких систем.

Среди существующих методов оценки характеристик текстов на естественном языке, применяемых поисковыми системами и системами автоматической обработки текста, были выделены и проанализированы следующие методы: проверка на естественность в соответствии с Законом Ципфа; алгоритм TF-IDF; алгоритм BM25 и BM25F; метод оценки скорости развития текстов (коэффициент новизны); метод оценки качества текста; метод оценки связности текста; метод оценки текста на переспам.

Анализ даже простых образцов бытовой речи приводит к довольно сложным комбинаторным структурам на уровне семантического представления. В исследованиях, посвященных обработке текстов на естественном языке, зачастую оказываются запутанными проблемы анализа семантики и создания языка смыслов и собственно перевода в обе стороны. Неясно, какой максимальный уровень структурированности может быть достигнут и каково происхождение неалгоритмизуемого остатка, является ли он следствием исторических случайностей и прихотливости естественного языка или есть более глубокие причины его существования.

Для решения описанной проблемы можно предложить следующий подход.

Удобно выбрать такую ограниченную область смыслов, чтобы ее семантическое представление имело наиболее простую и фиксированную структуру. Выберем в качестве этой области «натуральные числа», семантически представляя их последовательностями палочек: I, II, III, IIII, ... История ее представления средствами естественных языков относится к описательной лингвистике; одновременно она предоставляет ценные свидетельства о ранних стадиях математического мышления. Результаты автоматического анализа текста должны иметь вид смыслообразующих связей текста и не зависеть от способа, используемого при анализе [2].

Приведем наиболее распространенные метрики в области анализа текстов [3].

Проверка на естественность в соответствии с Законом Ципфа

Закон Ципфа представляет собой метод оценки естественности текста путем определения закономерности распределения частоты слов [2]. Естественность текста определяется близостью частотного распределения, полученного для исследуемого текста к эталонному распределению.

Эталонная кривая отражает следующую эмпирическую закономерность распределения частоты слов естественного языка: если все слова достаточно длинного текста упорядочить по убыванию частоты их использования, то частота *n*-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру *n*. Например, второе по используемости слово встречается примерно в два раза реже, чем первое, третье — в три раза реже, чем первое, и так далее.

Алгоритм TF-IDF

Алгоритм TF-IDF используется для расчета важности слова (веса) в документе. Этот показатель прямо пропорционален количеству вхождений слова в анализируемый текст и обратно пропорционален частоте употреблению этого слова в других доступных текстах антологии (в случае поисковых систем антологией считаются все доступные документы в сети Интернет).

TF (term frequency — частота слова) — отношение числа вхождения некоторого слова к общему количеству слов документа. Таким образом, оценивается важность слова t_i в пределах отдельного документа.

$$\mathrm{TF}(t,d) = \frac{n_i}{\sum_k n_k},$$

где есть число вхождений слова в документ, а в знаменателе — общее число слов в данном документе.

IDF (inverse document frequency — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции. Учёт IDF уменьшает вес широкоупотребительных слов. Для каждого уникального слова в пределах конкретной антологии существует только одно значение IDF.

$$\mathrm{IDF}(t,D) = log\,\frac{|D|}{|(d_i \supset t_i)|},$$

Где $|D|$ — количество документов в антологии;

$|(d_i \supset t_i)|$ — количество документов, в которых встречается t_i (когда $n_i \neq 0$).

Выбор основания логарифма в формуле не имеет значения, поскольку изменение основания приводит к изменению веса каждого слова на постоянный множитель, что не влияет на соотношение весов.

Таким образом, мера TF–IDF является произведением двух сомножителей:

$$\mathrm{TF\text{–}IDF}(t,d,D) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D).$$

Большой вес в TF–IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

Алгоритмы BM25 и BM25F

BM25 — поисковая функция на неупорядоченном множестве термов и множестве документов, которые она оценивает на основе встречаемости слов запроса в каждом документе, без учёта взаимоотношений между ними (например, близости).

Алгоритм BM25 пришел на смену TF–IDF, его суть заключается в оценке текста на странице, основываясь на количестве и месторасположении ключевых слов, без учета ссылок.

Алгоритм BM25F учитывает не только сам текст, но и его отдельные участки или зоны. К таким участкам относят тег Title, метатеги, заголовки и подзаголовки, околоссылочный текст.

В дополнение к известным метрикам авторами предлагаются следующие методы оценки качества текстов [4].

Метод оценки скорости развития текстов (коэффициент новизны)

Метод является развитием критерия, основанного на применении закона Ципфа, и позволяет определять скорость развития текстов (долю появления новых терминов в каждом последующем предложении по отношению к уже имеющемуся фрагменту текста) [2].

Метод оценки качества текста

Для повышения качества отделения научно-технических текстов от текстов рекламного характера предложен метод, основанный на оценке скорости уменьшения объема автоматически сформированного реферата текста на каждом шаге реферирования.

Рефератом является множество F_ε на каждом шаге $\varepsilon = 1, 2, \ldots, n$, которое формируется из предложений s исходного текста T по правилам $s \in \mathrm{F}_\varepsilon$, если $\rho(s) \geq \varepsilon$, где $\rho(s)$ — рейтинг предложения. Значение n определяется условием $|\mathrm{F}_n| = 0$.

В данном методе для определения рейтинга предложения используется понятие двусвязок слов, предложенное в работе [5]: синтаксическая связь k между двумя словами предложения.

Рейтингом предложения считается максимальный рейтинг элементов множества K_s двусвязок, входящих в предложение, рассчитываемый по формуле:

$$\rho(s) = \max_{k \in K_s} |L_k|, s \in L_k,$$

Где L_k — множество предложений, содержащих синтаксическую связь k между 2 словами.

Метод оценки связности текста

В качестве одного из вариантов оценки качества текстов предложен метод оценки связности семантического окружения, предложенного в [1].

Числовым показателем связности является:

$$Q = \frac{N_A}{N_W},$$

Где N_A – число связных областей, N_W=|W| – мощность множества W слов в тексте.

Метод оценки текста на переспам

Понятие переспама связано с превышением количества ключевых слов в документе и является следствием неправильной или излишне агрессивной поисковой оптимизации, также может сопровождаться частым выделением ключевых слов «жирным текстом», использованием спамных конструкций типа «книжный шкаф Москва» или «чайный сервис купить Воронеж» и т.п, переоптимизацией (употребление более одного раза ключевого слова).

Эмпирически определен порог переспама как превышение 3–4 вхождений каждого ключевого слова на 2 – 3 тысячи символов текста.

Анализ измеримых параметров текста

Кроме структурных и семантических параметров качества текста имеется также большой набор технологических параметров, оказывающих сильное влияние на качество работы автоматических систем анализа текстов на естественном языке.

1. *Степень нарушения правил построения синтаксических конструкций*

Выявление нарушения правил построения синтаксических конструкций имеет место при анализе текстов, представляющих собой результат живого общения или обсуждения (форумы, блоги, дневники). На увеличение значения данного параметра влияет отсутствие знаков препинания и использования заглавных букв в начале предложения. Влияет на уровень корректности синтаксического анализа и выделения структурных единиц текста.

2. *Степень нарушения правил орфографии (намеренное и случайное)*

Выявление нарушения правил орфографии имеет место при анализе текстов, представляющих собой результат живого общения или обсуждения (форумы, блоги, дневники). Большое значение данного параметра означает появление в анализируемом тексте неологизмов, по факту являющихся ошибочным написанием словарного слова. Вызывает проблемы с реализацией функций поиска.

3. *Степень использования не общеупотребимых сокращений*

Аналогично п.2.

4. *Степень неоднородности символьного набора в пределах каждой синтагмы (использование «чебурашки»)*

Неоднородность символьного набора в пределах каждой синтагмы возникает при замене некоторых символов в национальной кодировке на символы, имеющие похожие графемы, но содержащиеся в другой национальной кодировке.

Высокое значение параметра показывает, что имеет место прием использования «чебурашки» для обмана систем антиспама и удаления нецензурной лексики в текстах. Проблемы аналогичны п.2.

5. Наличие денотаций в тексте

Параметр, обозначающий количество определений терминов, имеющихся в тексте. С точки зрения этого показателя наиболее эффективными текстами для задачи обнаружения требуемой информации в тексте являются энциклопедии и словарные статьи.

6. Связность текста

Параметр текста, обозначающий степень связи элементов текста, определяющий целостность речевого сообщения и обусловленный авторским замыслом и особенностями языковой системы, стоящей за текстом. Формально может определяться через связность семантического окружения, построенного для текста.

7. Непротиворечивость

Параметр, показывающий отсутствие противоречия сведений, имеющихся во фрагменте документа по отношению к другим фрагментам документа. Значение параметра снижается при использовании неточных формулировок, плеоназмов и прочего языкового мусора.

8. Достоверность

Параметр, показывающий отсутствие противоречия сведений, имеющихся в документе по отношению к другим документам антологии и общим знаниям в данной предметной области. Значение параметра снижается при использовании неточных формулировок, неверных заимствований. Влияет на качество автоматически формируемого семантического окружения и онтологии при использовании такого текста.

9. Уникальность текста

Параметр, показывающий насколько входящие в его состав фразы уникальны в антологии. Чем выше заимствование из других источников, тем ниже будет этот показатель.

Следует отметить также качественные параметры, относящиеся к текстам, являющиеся необходимыми и достаточными для из автоматизированной обработки.

10. Доступность

Параметр, обозначающий возможность доступа к документу информационной системы. Является одним из необходимых условий функционирования информационных систем информационного поиска и обнаружения требуемой информации в коллекции документов.

11. Различимость текстов

Возможность технически различать несколько экземпляров документа. Нарушается, например, при невозможности анализа внутренней структуры формата (PDF, DJVU) или при использовании искусственных приемов подмены различимых метатегов для документов, представленных в виде графических файлов.

12. Идентифицируемость текстов

Возможность сформировать идентификатор, однозначно указывающий на конкретный документ в общей коллекции.

Рассмотренные параметры, относящиеся к текстам, можно разделить на две основные группы:

– Локальные параметры (параметр единичного текста) – совокупность характеристик, присущих каждому конкретному тексту (связность текста, соблюдение орфографии, непротиворечивость, наличие денотаций в тексте).

– Макропараметры — характеристики конкретного текста, которые имеют смысл при рассмотрении этого текста в совокупности с другими текстами (уникальность текста, доступность, различимость, достоверность).

Таким образом, в работе проведен анализ характеристик текстов на естественном языке, позволяющий использовать технически-реализуемые критерии для предварительной автоматизированной оценки текстов, подлежащих аналитическому мониторингу. Кроме того, рассмотренные характеристики могут использоваться в качестве базовых технических требований, предъявляемых к входным данным в информационных системах, работающих с текстами на естественном языке.

Литература:

1. В.В.Александров, С.В.Кулешов Аналитический мониторинг Internet контента. Инфологический подход — Качество. Инновации. Образование № 3, 2008 — с 68–70
2. В.В.Александров, А.В.Арсентьева, А.И.Семенков. Структурный анализ диалога. Препринт. Ленинград, 1983
3. Ю. Силин. Качество текста, способы оценки // Электронный ресурс — Доступ: http://inetmkt.ru/seo-optimizacia/kachestvo-teksta-sposobyi-otsenki (Дата 30.09.2015)
4. С.Н. Михайлов, С.В. Кулешов, А.А. Зайцева Критерии качества технических текстов в задачах аналитического мониторинга информационных ресурсов // Труды СПИИРАН. 2014. Вып. 6(37).
5. Кулешов С.В. Разработка автоматизированной системы семантического анализа и построения визуальных динамических глоссариев // Диссертация на соискание ученой степени кандидата технических наук. Санкт-Петербург. 2005. 100 с.

Создание системы управления рисками коммерческого банка на основе управленческого подхода

The Establishment of the Risk Management System of Commercial Bank Based on the Management Approach

Калимуллина О.В.,
Ассистент кафедры ИСТВБ,
Университет ИТМО, Санкт-Петербург

Ключевые слова: интегрированная система управления рисками, BSC, SWOT-анализ, метод аналитических сетей
Keywords: integrated risk management system, BSC, SWOT-analysis, method of analytical networks

Современный этап развития отечественной банковской системы характеризуется расширившимся спектром рисков, с которыми сталкиваются банки. Несмотря на то, что вопрос совершенствования систем риск-менеджмента кредитных организаций был всегда актуален, развитие методологии управления рисками было преимущественно ориентировано на регулятивный подход, при этом управленческому подходу к формированию интегрированной системы риск-менеджмента банка уделялось недостаточно внимания. Таким образом, очевидна необходимость дальнейшего совершенствования теоретических и методических подходов к созданию комплексной системы управления рисками с применением управленческого подхода.
Одним из инструментов управленческого подхода, позволяющих перевести риск-менеджмент из области тактических задач в стратегические, является BSC. На основе проведенного анализа были выявлены достоинства и недостатки BSC, для преодоления которых была обоснована целесообразность применения BSC совместно с методом аналитических сетей, что позволяет определить относительную важность составляющих BSC, присвоить веса показателям, а также