УДК 004.91 (574)

Гулнар АХМЕТЖАНОВА, *магистрант, Карагандинский экономический университет Казпотребсоюза, 100009, г. Караганда, ул. Академическая, 9,e-mail: Ags_83@mail.ru*

МЕТОДЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВЫХ РЕСУРСОВ Е-УНИВЕРСИТЕТА

В данном статье рассмотрены различные методы автоматической обработки текстовых ресурсов для е-университета. Цель статьи - изучение методов автоматической обработки текстовых ресурсов для е-университета. Разработка оптимальной модели автоматической обработки текстовых ресурсов для е-университета, обрабатывающие огромные объемы электронных текстов, в связи с развитием информационных сетей и повышением их мобильности при работе в многоязычной среде.

Ключевые слова: е-университет, методы экстраполяции, методы моделирования, Метод экспоненциального сглаживания, метод наименьших квадратов.

Гулнар АХМЕТЖАНОВА, *магистрант, Қазтұтынуодағы Қарағанды экономикалық университеті 100009, Қазақстан Республикасы, Қарағанды қ. Академическая көш 9, e-mail: Ags_83@mail.ru*

Е-УНИВЕРСИТЕТ МӘТІНДІК ҚҰРАЛДАРЫН АВТОМАТТЫ ӨҢДЕУ ӘДІСТЕРІ

Бұл мақалада е-университет үшін мәтіндік ақпаратты автоматты түрде өңдеуге арналған әр-түрлі әдістер қарастырылды.Мақаланың мақсаты: е-университет үшін мәтіндік ақпаратты автоматты түрде өңдеуге қажетті әдістерді қарастыру.Е-университет үшін мәтіндік ақпаратты автоматты түрде өңдеуге қажетті оптималды әдісті құрастыру, үлкен көлемді электронды мәтіндерді өңдеу, әр-түрлі көп тілді ортада жұмыс істеген кезде ақпараттық желілерді дамытуға және олардың мобильділігін арттыруға байланысты.

Түйінді сөздер: электрондық университеттер, экстраполяция әдістері, модельдеу әдістері, экспоненталды тегістеу әдісі, кіші квадраттар әдісі.

Gulnar AKHMETZHANOVA, *master's stuent Karaganda economic University of Kazpotrebsoyuz, 100009, Karaganda, Akademicheskaya str 9 e-mail: vip_nurbek@mail.ru*

METHODS OF AUTOMATIC TREATMENT OF TEXT RESOURCES E-UNIVERSITY

In this article, various methods of automatic processing of text resources for e-university are considered. The purpose of the article is to study methods of automatic processing of text resources for e-university. Development of an optimal model for the automatic processing of text resources for e-university, processing huge amounts of electronic texts, in connection with the development of information networks and increasing their mobility when working in a multilingual environment.

Keywords: e-university, extrapolation methods, modeling methods, exponential smoothing method, least-squares method.

При создании автоматической обработки текстовых ресурсов e-University, то есть автоматизированной информационной системы, которая представляет собой набор программ, интегрированных в единую информационную среду и позволяющих автоматизировать основные процессы обработки информации, возникают проблемы автоматической обработки текстовой информации на первый план. Это определяется тем фактом, что естественный язык является не только инструментом мышления, но и универсальным средством коммуникации - средством восприятия, накопления, хранения, обработки и передачи информации [1]. Более того, естественный язык является универсальным средством описания реальности и общения с компьютерной системой.

В современном обществе, когда почти каждый может быть пользователем, проблема взаимодействия человека с технологией на естественном языке стала важной практической задачей. Сегодня автоматическая обработка естественного языка (включая текстовую и другую информацию) - это быстро развивающаяся область исследований и коммерческого развития, направленная на развитие систем промышленной обработки информации, которые должны быстро и эффективно обрабатывать в реальном времени огромные информационные потоки, циркулирующие в информационных системах с новейшими технологиями.

Автоматическая обработка текстовых ресурсов включает в себя решение многих задач, которые

можно условно разделить на два уровня. Задачи высокого уровня представлены задачами распознавания речи, форматирования текста, генерации документов, машинного перевода, поиска информации, т.д. Задачи самого низкого уровня включают грамматический анализ, удаление смысловых значений, коррекцию орфографического и синтаксического анализа, т.е. Задачи лингвистической обработки тестовой информации [2].

Цель построения текстовых моделей отличается от построения других модели, чтобы сделать причинные выводы - основная цель большинства моделей технической науки. Обычный совет для причинной построение модели вывода заключается в том, что важно включить все соответствующие функции источника данных процесса - либо в структуре модели.

Круг задач значительно расширился и в целом охватывает всю индустрию развития и поддержки информационной технологии в управления и обработки текстовых данныхв том числе при обработке данных в е-университете. Одним из важнейших особенной при создание автоматизированных информационных систем "е-университета", является проблема при решении указанных задач состоит в необходимости обрабатывать неструктурированные тексты.

Информационные системы ориентированы на поиск алгоритмов решение интеллектуальных задач, единый типовой алгоритм их автоматической обработки создать не удается, поскольку конкретный вид алгоритма, в первую очередь, определяется строем языка [3].

Существует два подхода к решению проблем при автоматизации информационной системы электронного университета при обработке текстовых данных [4]:

- в первом развивается в рамках искусственного интеллекта и называется интеллектуальным подходом. Он основан на предположении, что компьютерная система для успешной обработки информационных данных должна иметь возможность привлекать огромные ресурсы знаний о мире и делать логические выводы на основе этих знаний.

- во втором подходе инженерно-лингвистический подход был сформирован в компьютерной лингвистике и основан на концепции воспроизводящей инженерно-лингвистической модели. Эта модель подразумевает изучение всех лингвистических данных: от общего приближенного, основанного на изучении больших объемов реальных текстов информационных данных, описывающих феномен языка, посредством его структурного формального описания для алгоритмического воспроизведения описанного явления в рамках системы автоматической обработки информационных данных [2].

Информационные системы, откликаясь на запросы пользователей, развивались в направлении усложнения класса решаемых задач, систем и технологий. В настоящее время широкое применение находят интеллектуальные информационные системы, выполняемые более полные функциональные задачи для сложных систем.

Эти новые типы информационных систем воплощают знания, которые позволяют им проявлять интеллектуальное поведение. Автоматическая обработка текстовых ресурсов е-университета с системами в таких задачах, как решение проблем и поиск и манипулирование большим разнообразием мультимедийной информации и знаний. Эти системы выполняют такие задачи, как ориентированный на знания вывод, чтобы обнаружить знания из очень больших коллекций данных и обеспечить совместную поддержку пользователям в сложном анализе данных. Текстовые ресурсы е-университета также занимается поиском, доступом, извлечением, хранением и обработкой больших тестовых информации и знаний, а также интеграцией информации и знаний из множества гетерогенных источников.

Автоматическая обработка текстовых документации е-университете имеет важную функцию - составление отчетности и предоставление ее компетентным органам. Автоматизация делает возможным автоматически подсчитать людей по разным категориям: в зависимости от возраста, пола, национальности. Собранные данные могут сыграть не только информационную, но и предупреждающую роль для университета. Информация о контингенте студентов, о численности студентов на данный период может служить базисом для прогнозов на будущие периоды. Это позволит исходя из полученных данных откорректировать некоторые действия, для предотвращения прогноза, выявить факторы, влияющие на развитие событий и по возможности устранить или же сгладить нежелательное влияние информационных данных, при обработке текстовых информаций о студенте.

Следует отметить, что данная информация позволит:

- скорректировать политику в области профессиональной ориентации абитуриентов, с целью ориентируются на тщательное изучение экономического объекта и его моделирование;

- спланировать учебную нагрузку, количество учебного материала на различных языках, организовать соответственный досуг для всех категорий студентов;

- выявить причины оттока студентов на различных специальностях, провести работу по поднятию популярности специальности;

- определить причину оттока студентов в другие вузы с целью устранить ее и улучшить условия обучения в вузе.

Следовательно это база основной текстовой информаций в е-университет которая необходимо для создание различных видов запроса, при необходимости для отчетности, для обработки данных информативных вседений о студентах с использванием методов прогназирование коректных данных.

Базы знаний систем, основанных на прецедентах (Case-based reasoning) содержат описания разнообразных ситуаций. Поиск решения проблемы сводится к поиску по аналогии (абдуктивному выводу от частного к частному) подходящей ситуации и включает следующие шаги [4].

1. Получение подробной информации о текущей проблеме.

2. Сопоставление полученной информации со значением признаков прецедентов из базы знаний.

3. Выбор прецедента из базы знаний, наиболее близкого к рассматриваемой проблеме.

4. В случае необходимости выполняется адаптация выбранного прецедента к текущей проблеме.

5. Проверка корректности каждого полученного решения.

6. Занесение информации о полученном решении в базу знаний.

В отличие от индуктивных систем допускается нечеткий поиск с получением множества допустимых альтернатив в задачах, каждая из которых оценивается некоторым коэффициентом уверенности. Обучение системы сводится к запоминанию каждой новой обработанной ситуации с принятыми решениями в базе прецедентов, и выполнение последовательностей алгоритмов. Данные системы применяются как системы распространения знаний с расширенными возможностями или как системы контекстной помощи, относящиеся к системе знаний е-университета.

В информационных систем является постоянно развиваемая модель проблемной области, поддерживаемая в специальной базе знаний - репозитории, на основе которого осуществляется генерация или конфигурация программного обеспечения. Таким образом, проектирование и адаптация системы сводится, прежде всего, к построению модели проблемной области и ее своевременной корректировке. При проектировании адаптивной системы обычно используется два подхода: оригинальное или типовое проектирование. Первый подход предполагает разработку ИИС с "чистого листа" в соответствии с требованиями экономического объекта. Он реализуется с использованием систем автоматизированного проектирования информационных систем или CASE технологий. Согласно этой технологии каждый раз при изменении проблемной области выполняется генерация программного обеспечения. При втором подходе выполняется адаптация типовых разработок к особенностям экономического объекта. Он предполагает использование систем компонентного (сборочного) проектирования.

Основное внимание здесь уделяется изучению аппроксимирующих, вероятностных и логических механизмов получения общих выводов из частных утверждений в процессе прогнозирования данных обработанной информации в системе знаний, в том числе тестовой информаций.

Можно выделить пять этапов при процессе прогнозирования: сбор данных, редукция данных, построение модели и ее оценка, экстраполяция выбранной модели, оценка полученного прогноза. Привлекает возможность получить систему, настраивающуюся на сколь угодно сложное поведение и адекватное решаемой задаче.

Еще одним достоинством в случае аппаратной реализации сбор данных, подразумевает получение корректных данных и верификация этих данных. Данный этап зачастую является наиболее сомнительной частью всего процесса прогнозирования и сложным для проверки. Когда необходимо получить в е-университете основные определенные данные о студентах, то их сбор и проверка как правило сопровождается множеством различных проблем связанных с запросами в техническом уровне.

В случае программной реализации структурная редукция данных, важна для прогнозирования, в виду того, что в процессе сбора данных может быть собрано либо слишком много данных, либо слишком мало в сведений о студентах. В долговременной памяти хранятся не столько факты и данные, сколько объекты и связи между ними. В настоящее время широкое применение находят интеллектуальные информационные системы. Могут быть и данные, не имеющие непосредственного отношения к рассматриваемой задаче, а это будет способствовать снижению точности прогноза. Большие объемы данных постоянно записываются в кратковременную память, и мы непрерывно анализирует и фильтруем получаемую информацию.

Но для количественных методов анализа текста не используется. Включая более реалистичные функции в количественные модели не обязательно переводят на улучшенный метод и сокращение используемых допущений не может означать более продуктивный анализ. Скорее, тонкости применение методов к любому набору данных означает, что модели, которые менее сложны в использовании языка может обеспечить более полезный анализ текстов.

То, что все автоматизированные методы основаны на неправильных моделях языка, также подразумевает, что модели должны оцениваться на основе их способности выполнять какую-то полезную социальную научную задачу. Как мы объясняем ниже, акцент в оценках должен быть сделан на том, чтобы помочь исследователям назначить документы в предопределенные категории, открывать новые и полезные схемы категоризации текстов, или в измерении теоретически значимых величин из больших наборов текста. Альтернативная модель оценки, которые полагаются на модель, соответствуют или предсказывают содержание новых текстов, могут по существу выбирать.

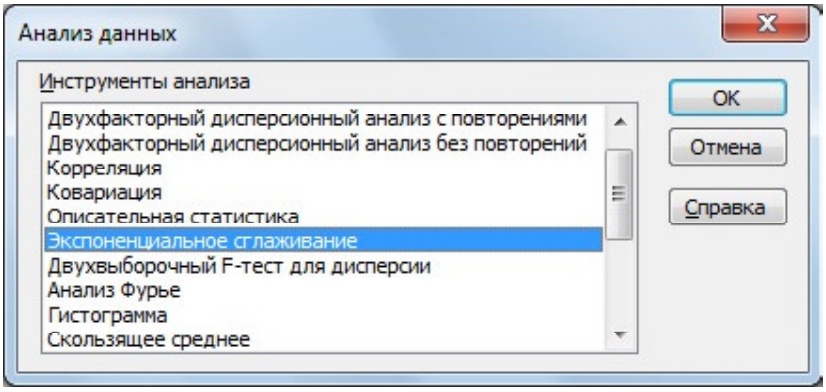
Автоматизированные методы анализа контента продемонстрировали эффективность во многих проблемах. Однако эти методы не исключают необходимости тщательного изучения исследователями или устранения необходимости в чтении текстов

Экстраполяция выбранной модели, которая в настоящее время разрабатывает новое направление в построении интеллектуальных систем, допускает нечеткий поиск с набором приемлемых альтернатив, каждый из которых оценивается с помощью некоторого коэффициента достоверности. Обучение системы сводится к запоминанию каждой новой обработанной ситуации с принятыми решениями в базе прецедентов, предполагает фактическое получение объявленного прогноза, аналитически модель линейного тренда позволяет получать полную информацию.

Под оценкой полученного прогноза понимается программная система, выполняющая действия, аналогичные тем, которые выполняются экспертом в какой-либо прикладной предметной области, делая определенные выводы в ходе выдачи рекомендаций и рекомендаций по перемещению данных, заключается в сравнении вычисленных значений с наблюдаемые значения. После выбора прогнозной модели прогноз выполняется за указанные периоды, а полученные

результаты сравниваются с известными наблюдаемыми значениями. Прогнозирование - это предсказание будущего развития систем, основанных на прошлом и настоящем поведении; системы содержат блоки для обработки статистики, принятия решений на основе неполной информации и создания альтернативных путей развития системы.

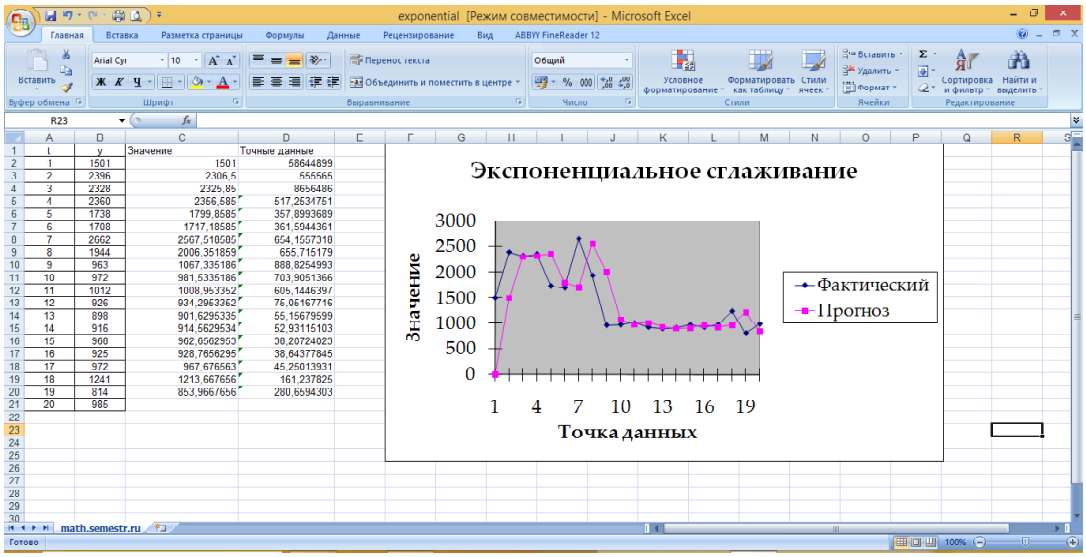
Таким образом, прогнозирование может помочь спланировать некоторые события, решить многие проблемы. Метод прогнозирования используется для определения исследований объектов прогнозирования, направленных на разработку прогнозов. Этот метод позволяет определить набор специальных методов и правил для разработки конкретных прогнозов. Получение прогноза - это математическая или логическая операция, направленная на получение конкретных результатов в процессе разработки прогнозов. Рассмотрены различные подходы, используемые при проектировании и разработке интеллектуальных систем и технологий в комплексе, а также рассмотрены тенденции развития поисковых систем прогнозирования. Прогноз поиска является условным продолжением наблюдаемых тенденций исследуемого явления или процесса, существующих решений, которые могут существенно изменить возникающие тенденции. Целью прогноза поиска является то, что проблемы могут возникнуть в будущем, сохраняя при этом существующую тенденцию к методам автоматической обработки текстовых данных. Проектирование системы сводятся, прежде всего, к построению модели проблемной области и ее своевременной корректировке текстовых данных. Данный метод можно рассчитать и определить с помощью пакета прикладной программы Microsoft Office Excel 2013 указана на рисунке 1 и основные расчеты даны на рисунке 2.



Примечание - Составлено автором

Рис 1. Выбор метода экспоненциального сглаживание

Методы планирования решения помогают специалистам выбрать и сформировать оптимальное решение в конкретной ситуации. Поддержка принятия решений позволяет определит процедуру, обеспечивающее принятие решения, необходимой информацией и рекомендациями, облегчающими процесс принятия решения задачи экспоненциальным методом (рисунок 2). Он реализуется с использованием пакета прикладных программ Microsoft Office Excel более развернутом виде, в расчете указаны обработанные средние значение тестовых операций и точные данные обрабатываемых операций в прогнозируемый период времени. Учитывая неформализованность решаемых задач и эвристический, личностный характер используемых знаний, есть возможность непосредственно взаимодействовать с информационной системой.



Примечание - Составлено автором по данным [2, 3]

Рис 2. Расчеты данных экспоненциального сглаживание

При сглаживании временного ряда фактических данных скользящими средними в расчетах участвуют все уровни ряда это указана на рисунке 2. Чем шире интервал сглаживания, тем более плавным получается тренд. Сглаженный ряд короче первоначального на (n-1) тренд наблюдений, где n - величина интервала сглаживания тренда. Согласно этой технологии каждый раз при изменении проблемной области выполняется генерация программного обеспечения. Он предполагает использование систем компонентного (сборочного) проектирования тестовых информаций в обрабатываемой среде.

При больших значениях n колеблемость сглаженного ряда значительно снижается в тренде. Одновременно заметно сокращается количество наблюдений тренда, механизм логического вывода - часть сглаживании, реализующая анализ поступающей в экспоненциального сглаживание и имеющейся в ней информации и формирование (вывод) на ее основе новых заключений сглаживании в ответ на запрос к системе, что создает трудности. Компонент приобретения знаний предназначен для обеспечения работы инженера знаний по поддержанию модели знаний, адекватной реальной предметной области.

Выбор интервала для обработки информации для определения конкретных фактов использует общие правила. Изучение на основе подобия является индуктивным процессом, и доказательство теорем сглаживается, поскольку оно опирается на известные аксиомы и уже доказанные теоремы, сглаживание зависит от целей исследования. В то же время следует руководствоваться периодом времени, в течение которого происходит действие, и, следовательно, устранением влияния случайных факторов. Он объясняет, как система получила решение проблемы (или почему она приняла такое решение) и какие знания она использовала при ее использовании, что облегчает эксперту проверку системы и повышает доверие пользователя к результату.

Простой способ сглаживания временных рядов - просто под выбор ряда; например, заменить ежедневные данные еженедельными данными. Хотя этот метод подвержен ошибкам выборки, он может уменьшить дисперсию временных рядов. Например, при геометрической модели броуновского движения она уменьшает волатильность, пропорциональную отношению квадратных корней интервалов.

Данный метод используется при краткосрочном прогнозировании сглаживания временного ряда в потоке информационных данных. Его рабочая формула (1):

$$y_{t+1} = m_{t-1} + \frac{1}{n} x(y_t - y_{t-1}), если\ n = 3, \qquad (1)$$

где t + 1 - прогнозный период;

t - период, предшествующий прогнозному периоду (год, месяц и т.д.);

y_{t+1} - прогнозируемый показатель;

m_{t-1} - скользящая средняя за два периода до прогнозного;

n - число уровней, входящих в интервал сглаживания;

y_t - фактическое значение исследуемого явления за предшествующий период;

y_{t-1} - фактическое значение исследуемого явления за два периода, предшествующих прогнозному.

Модель предполагает, что данные текут вокруг довольно стабильного среднего (нет тенденции или постоянной динамики роста) в потоке информационных данных, в том числе приемлемых для тестовой информации. Экспериментальные результаты показывают, что использование более сложных представлений не улучшает результаты, поскольку для получения преимуществ дополнительной информации следует разработать более сложные методы. В этом случае наличие меньшего количества очень информативных признаков может дать результаты.

Преимущество этой модели состоит в том, что параметры инициализируются только один раз. Следовательно, как только параметры были установлены, предсказание может продолжаться без какой-либо задержки в пересчете параметров. Другим преимуществом является то, что прошлые данные не нужно помнить.

В заключение следует отметить автоматической обработки текстовых ресурсов для е-университета и информационных технологий, данный подхода ориентируются на тщательное изучение экономического объекта и его моделирование, которые делают реальным проведение исследований с помощью методов сглаживания временных рядов:

- доступность относительно основных оперативных данных из базы, автоматической обработки текстовых ресурсов для е-университета с достаточным быстродействием и объемом памяти для анализа большого количества данных;

- существование больших корпусов лингвистических и лексических данных (словари, тезаурусы, тексты) для обучения и тестирования систем в автоматической обработке текстовых данных;

- возрастание спроса на коммерческие системы, обрабатывающие огромные объемы электронных текстов, в связи с развитием информационных сетей и повышением их мобильности при работе в многоязычной среде;

- появление информационных технологий, способных при решении определенного класса задач методов сглаживания временных рядов достигать довольно высокого уровня точности при работе с реальными данными.

Этот подход позволяет вычислить относительное правдоподобие конкурирующих гипотез исходя из различных методов прогнозирования, точнее метод сглаживания временного ряда.

Ориентация на природу методов сглаживания временных рядов позволяет преодолевать сбойные и тупиковые ситуации, возникающие на том или ином уровне обработки текста автоматизированной системы е-университета.

Список литературы:

1. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска // Ученые записки Казанского Государственного Университета. Серия Физико-математические науки. 2007. т.149. книга 2. С.49-72.

2. Алексеев А.А., Лукашевич Н.В. Комбинирование признаков для извлечения тематических цепочек в новостном кластере // Труды Института системного программирования РАН. 2012, Т. 23. С. 257-276. 15.

3. Лукашевич Н.В. Отношения часть-целое: теория и практика // Нейрокомпьютеры: разработка, применение. 2013. N1. С. 7-12.

4. Поллак, Г.А. Интеллектуальные информационные системы: учебное пособие / Г.А. Поллак - Челябинск: Издательский центр ЮУрГУ, 2011. - 141 с.