Методы автоматической обработки тем сообщений в потоках новостных сообщений

Зевайкин А.Н.
ИКСИ

Постановка задачи

Объект: потоки новостных сообщений

Цель: автоматическое выделение и представление актуальных тем в потоке новостей.

Актуальность задачи

Актуальность задачи обосновывается:

- С одной стороны, потребностью получать в реальном масштабе времени наиболее полные и точные сведения об окружающей обстановке.
- С другой стороны, идет постоянный рост объемов доступной текстовой информации, которую уже невозможно обрабатывать ручными методами.

Отличие от существующих систем

Многие существующие системы обработки текстовых данных способны работать с уже известными, заранее определенными понятиями, такими как поисковый запрос и образ рубрики.

Но эти системы не способны в полной мере оперировать с новыми неизвестными понятиями, такими, как только что произошедшее событие.

Типичный день аналитика:

1. Обойти все интересующие новостные сайты
2. Выделить для себя самые актуальные темы
3. Создать дайджест актуальных новостей

Используемые понятия

- Сообщение - единичный текстовый документ, поступающий из некоторого источника.
- Тема - «тема - предмет описания, изображения, исследования, выступления, дискуссии». В новостных системах тема описывается множеством сообщений, связанных между собой общим событием.

Модель темы

Тема – абстрактное понятие, описываемое однородной группой похожих, в определенном смысле, сообщений.

Ограничение автоматизированных систем

Любая автоматизированная система не способна однозначно выделить темы, она может лишь описать ее множеством сообщений, сама тема складывается в голове у пользователя системы после ознакомления с данным множеством сообщений.

Методы автоматической обработки тем

- Выделение тем
- Ранжирование тем
- Представление тем

Методы автоматической обработки тем

- Выделение тем
  - Кластеризация сообщений с использованием структуры текста
- Ранжирование тем
  - Введение единого ранга «актуальность» и ранжирования по нему
- Представление тем
  - Аннотирование тем
  - Аннотирование сообщений
  - Ранжирование сообщений

Методы автоматической обработки тем

- Выделение тем
  - Кластеризация сообщений с использованием структуры текста
- Ранжирование тем
  - Введение единого ранга «актуальность» и ранжирования по нему
- Представление тем
  - Аннотирование тем
  - Аннотирование сообщений
  - Ранжирование сообщений

Кластеризация текстовых сообщений

Целью кластеризации сообщений является автоматическое выявление групп лексически похожих сообщений среди заданного фиксированного множества сообщений.

Формальная модель текста

Тексты представляются векторами в элементарной теоретико-множественной модели. В качестве информационных признаков выбраны простые термины, приведенные к нормальной форме с помощью морфоанализа. Для снижения размерности используется селекция и трансформация признаков.

Использование структуры текста

Авторы сообщений вносят дополнительную смысловую структуру в текст, разбивая его на абзацы – части текста, характеризующиеся единством и относительной законченностью содержания.

Данное разбиение позволяет выделить отдельные мысли в тексте и использовать это для улучшения кластерного анализа.

Метод кластерного анализа текстов с разбиением на абзацы

1. Выделение абзацев
2. Кластерный анализ абзацев
3. Переход от групп абзацев к группам документов

Эффективность кластерного анализа текстов с разбиением на абзацы

Применение разбиения на абзацы позволяет уменьшить относительную ошибку кластеризации в 2 раза.

Методы автоматической обработки тем

- Выделение тем
  - Кластеризация сообщений с использованием структуры текста
- Ранжирование тем
  - Введение единого ранга «актуальность» и ранжирования по нему
- Представление тем
  - Аннотирование тем
  - Аннотирование сообщений
  - Ранжирование сообщений

Понятие «актуальности»

Согласно БСЭ, «Актуальность - важность, значительность чего-либо в настоящее время, современность, злободневность».

Понятие «актуальности темы»

Тема является актуальной, если она обладает следующими признаками:

1. Тема - новая по времени, то есть описывается свежими сообщениями.
2. Тема - важная, то есть описывается сообщениями, отражающими интерес пользователей и источников к данной теме.

Основные факторы актуальности тем

1. Время
2. Важность
    1. для пользователя
    2. для источников

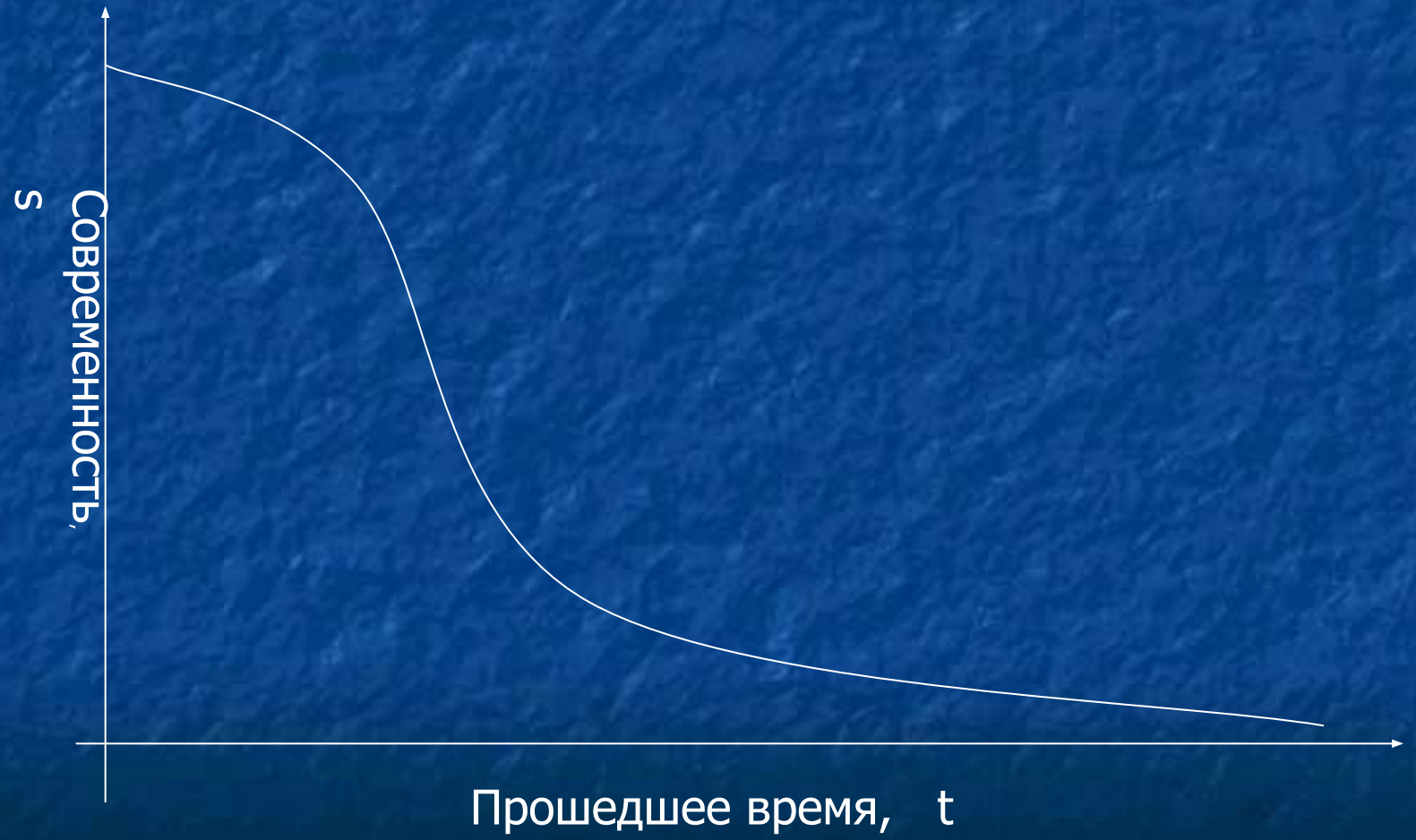
Основные факторы актуальности тем

1. Время
2. Важность
    1. для пользователя
    2. для источников

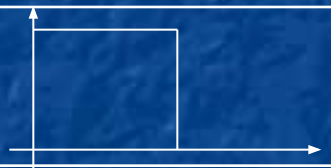
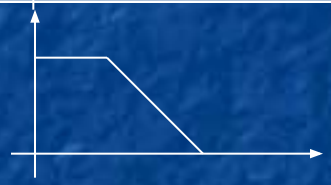
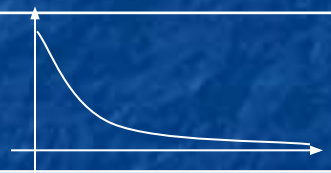
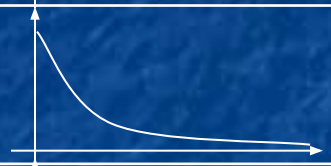
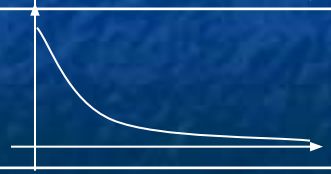
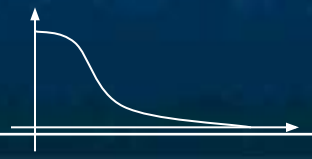
Ранжирование тем по времени

Сначала вычисляется среднее или последнее время сообщений в группе, затем время нужно подставить в функцию старения s(t).

Вид функции старения



Примеры функций современности

Название	Определение	График
Ступенчатая пороговая	s=1, при 0<t<T, s=0, при t≥T	
Линейная пороговая	s=1, при 0<t<T1, s=kt+b, при T1≤t≤T2, s=0, при t>T2	
Сигмоидная	s=2-2/(1+exp(-t))	
Гиперболический тангенс	s=1-th(t)=1-(exp(t)-exp(-t))/(exp(t)+exp(-t))	
Арктангенс	s=1-2*arctg(t)/п	
Гауссиана	$s=\exp(-k*x^2)$	

Основные факторы актуальности тем

1. Время
2. Важность
    1. для пользователя
    2. для источников

Ранжирование тем по важности для пользователя

Важность для пользователя мы можем рассчитать по количеству чтений сообщений из данной темы. Чем больше сообщений, тем более тема интересна пользователям.

Группы пользователей

При большом количестве пользователей имеет смысл разделение пользователей на группы по интересам.

Пользователь будет относиться к одной из групп, и ранг тем по важности для пользователя будет учитывать интересы группы.

Ранг по важности для пользователя с учетом групп

Ранг темы по важности для пользователя с учетом групп будет равен:

$$R_{user} = a_0 N_{read0} + a_1 N_{read1}, \quad a_0 < a_1$$

где N_{read0}, N_{read1} – число чтений пользователей, соответственно, из «чужих» групп и «своей» группы, a_0, a_1 – коэффициент, соответственно, «чужих» и «своей» группы.

Преимущества применения групп пользователей

Ранг тем будет динамически изменяться в зависимости от группы пользователя, и ранг будет выше у тех сообщений, которые больше интересны пользователям «своей» группы.

Основные факторы актуальности тем

1. Время
2. Важность
    1. для пользователя
    2. для источников

Ранжирование событий по важности для СМИ

Количество сообщений в группе отображает общий интерес новостных источников к данному событию. Чем больше пишут о данном событии, тем более оно интересно.

Ранжирование событий по важности для СМИ

Возможен более сложный вариант учета сообщений от источников: суммирование количества сообщений от данного источника умноженных на вес источника. Этим способом мы сможем отбросить излишние цитирования и сомнительные новости.

Ранжирование событий по важности для СМИ

Остается неучтенным вариант, когда один источник, пусть даже с малым весом, будет посылать большое количество сомнительных новостей на одну тему, в этом случае данная тематика подняться выше других, что неправильно.

Следует учитывать и долю источников, пишущих о данной теме, чем больше, тем лучше.

Ранжирование событий по важности для СМИ

Формула ранга важности для СМИ будет иметь следующий вид:

$$R_{smi} = \log \frac{i}{i-k+1} \sum_i v_i n_i$$

, где i – число источников,

k- число источников, пишущих на данную тему,

v_i - вес источника,

n_i - количество сообщений из данного источника на данную тему.

Формула актуальности темы

$$R_{full}=F_{full}(R_{time}, R_{user}, R_{smi})$$

Простейшая формула актуальности темы

$$F_{full} = R_{time}^{a_{time}} * R_{user}^{a_{user}} * R_{smi}^{a_{smi}}$$

, где a_{time}, a_{user}, a_{smi} – соответствующие коэффициенты рангов по времени, важности, задаваемые пользователем.

Формула актуальности темы

Более гибкий и сложный вариант – многокритериальное ранжирование.

$$R_{full} = y_{full} = (y_{time}\, y_{user}\, y_{smi})^{1/3} =$$

$$= ((a_{time} R_{time}^3 + b_{time} R_{time}^2 + c_{time} R_{time} + d_{time}) *$$

$$* (a_{user} R_{user}^3 + b_{user} R_{user}^2 + c_{user} R_{user} + d_{user}) *$$

$$* (a_{smi} R_{smi}^3 + b_{smi} R_{smi}^2 + c_{sme} R_{smi} + d_{smi}))^{1/3}$$

Методы автоматической обработки тем

- Выделение тем
  - Кластеризация сообщений с использованием структуры текста
- Ранжирование тем
  - Введение единого ранга «актуальность» и ранжирования по нему
- Представление тем
  - Аннотирование тем
  - Аннотирование сообщений
  - Ранжирование сообщений

Аннотирование тем

Предлагается использование результатов кластерного анализа с разбиением на абзацы для реферирования полученных тем. Выделяются абзацы, ближайшие к центру кластера, содержание каждого такого абзаца будет наиболее близко к теме соответствующего кластера.

Полученные абзацы представляют собой законченные смысловые блоки текста, наиболее близкие к данной теме, то есть реферат темы.

Методы автоматической обработки тем

- Выделение тем
  - Кластеризация сообщений с использованием структуры текста
- Ранжирование тем
  - Введение единого ранга «актуальность» и ранжирования по нему
- Представление тем
  - Аннотирование тем
  - Аннотирование сообщений
  - Ранжирование сообщений

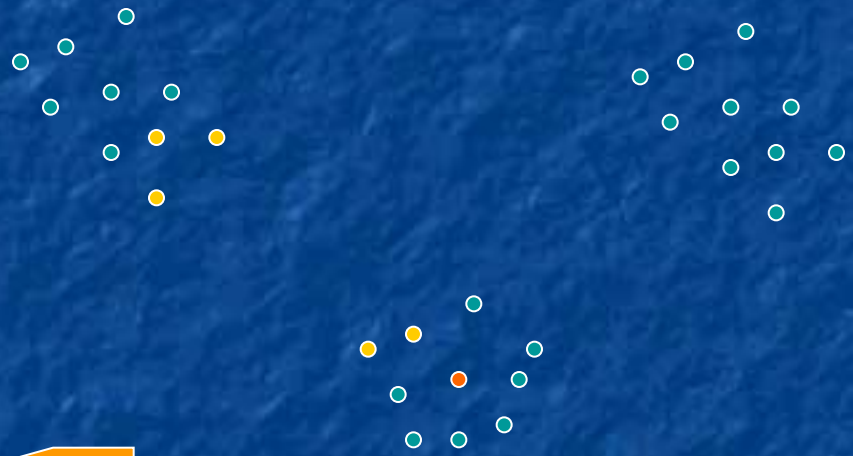
Аннотирование сообщений

Для каждого сообщения в теме (кластере) можно найти один или несколько абзацев, которые будет наиболее близки к центру данного кластера.

Данные абзацы будет являться выдержкой из текста, которая наиболее близка по содержимому к выбранной теме, то есть кратким описанием сообщения как элемента темы.

Наглядное представление метода аннотирования

Кластер, описывающий тему



Абзацы одного сообщения

Центральный абзац кластера

Методы автоматической обработки тем

- Выделение тем
  - Кластеризация сообщений с использованием структуры текста
- Ранжирование тем
  - Введение единого ранга «актуальность» и ранжирования по нему
- Представление тем
  - Аннотирование тем
  - Аннотирование сообщений
  - Ранжирование сообщений

Пример аннотирования

- ПО ДАННЫМ ПАРАЛЛЕЛЬНОГО ПОДСЧЕТА 67,3%% БЮЛЛЕТЕНЕЙ В ШТАБЕ ЯНУКОВИЧА, ЗА ПРЕМЬЕРА ПРОГОЛОСОВАЛИ 50,54

Как заявила журналистам представитель штаба Януковича Раиса Богатырева, после обработки 67,3%% бюллетеней центром параллельного подсчета голосов при штабе за Януковича проголосовали 50,54%%, за Ющенко - 45,53%%.

- НАБЛЮДАТЕЛИ ОТ СНГ НЕ ЗАФИКСИРОВАЛИ СЕРЬЕЗНЫХ НАРУШЕНИЙ НА ВЫБОРАХ ПРЕЗИДЕНТА УКРАИНЫ

В частности, в Одессе, Львове, Киеве наблюдалось несвоевременное открытие избирательных участков, уточнил собеседник агентства. Также, по его словам, во Львове, Херсонской области и Луцке на отдельных избирательных участках в кабины для голосования заходили сразу несколько человек.

Пример аннотирования системы «Яндекс Новости»

- Украина: взлом сейфа и гонки по вертикали 11:21 Правда.ru

Со всех уголков Украины продолжает поступать информация о нарушениях и ...

... списков и бюллетеней только в 14 часов в воскресенье, сообщает МВД Украины.

- Оппозиция на улице, в ЦИКе перерыв 11:05 РБК

ЦИК Украины объявил перерыв в подсчете голосов до 15 часов.

... обработки Центральной избирательной комиссией Украины 75,26% протоколов стало ...

Ранжирование сообщений в выбранном событии

- Ранжирование сообщений по времени
- Ранжирование сообщений по содержанию

Ранжирование сообщений по времени

Использует подобную функцию, как и в случае ранжирования событий.

Ранжирование сообщений по содержанию

- Близость сообщения к центру группы.
- Процент абзацев сообщения, наиболее близких тематике события.

Формула ранга сообщения

Подход к вычислению итогового ранга аналогичен подходу вычисления актуальности темы.

Простой случай – произведение рангов,

Сложный случай – многокритериальное ранжирование.

Заключение

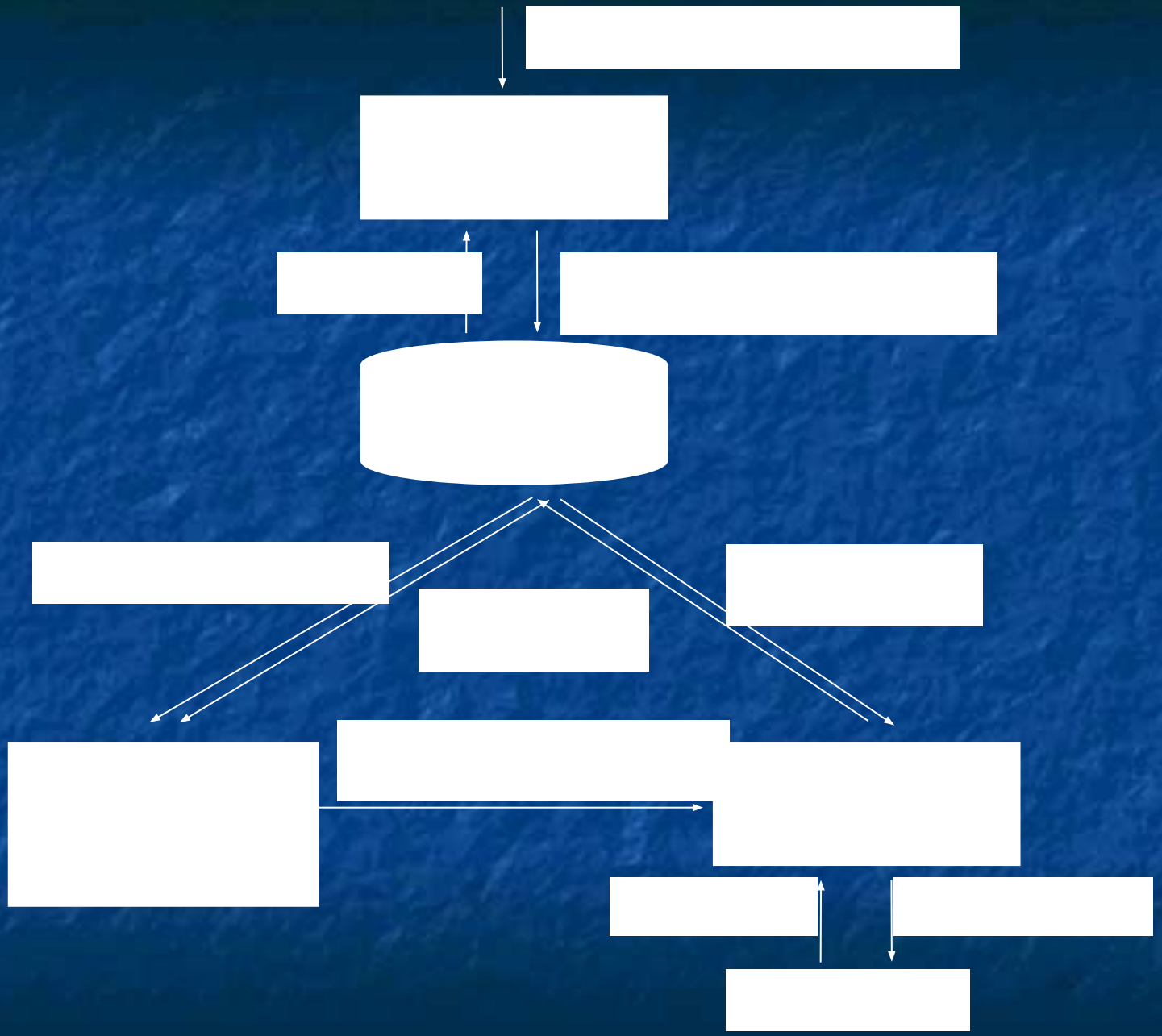
Рассмотренная задача автоматического определения актуальности сообщения отличается от уже существующих задач аналитической обработки текстов более широким подходом к анализу новых сообщений.

Предлагается анализировать не только содержимое текстов, но большое число других факторов, что позволяет в итоге более точно ранжировать сообщения по степени актуальности их для аналитика, обращать внимание на наиболее важные и свежие сообщения и пропускать ненужные.

Новизна исследования

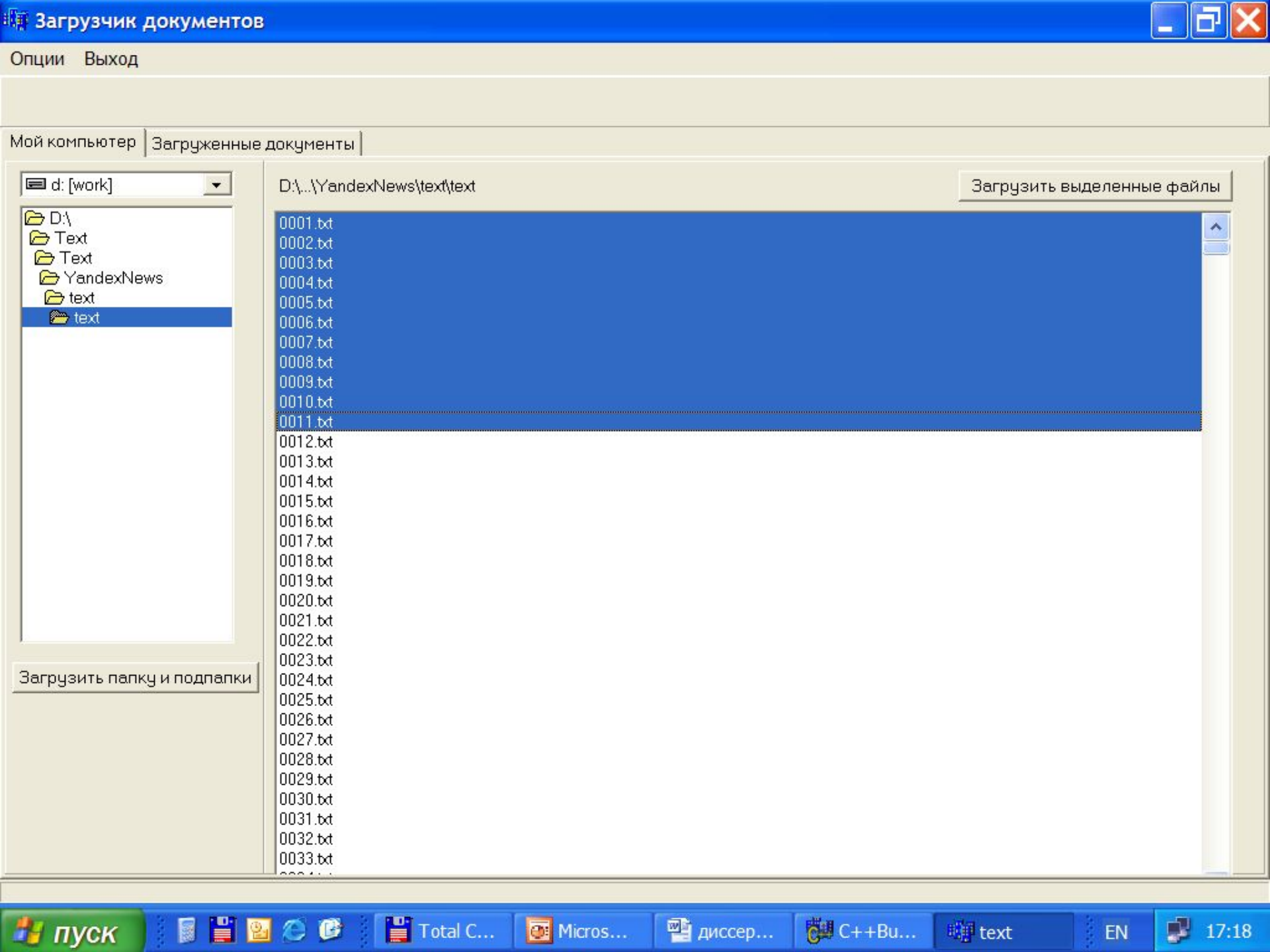
Разработаны:

- Метод кластерного анализа текстовых сообщений с использованием структуры текста
- Метод ранжирования тем сообщений
- Метод наглядного представления тем и сообщений



Подсистема предварительной обработки

- Лингвистическая обработка, формирование векторного представления.
- Формирование паспорта сообщения.


Загрузчик документов
Опции Выход
Мой компьютер Загруженные документы
d: [work]
D:\
Text
Text
YandexNews
text
text
D:\...\YandexNews\text\text
Загрузить выделенные файлы
0001.txt
0002.txt
0003.txt
0004.txt
0005.txt
0006.txt
0007.txt
0008.txt
0009.txt
0010.txt
0011.txt
0012.txt
0013.txt
0014.txt
0015.txt
0016.txt
0017.txt
0018.txt
0019.txt
0020.txt
0021.txt
0022.txt
0023.txt
0024.txt
0025.txt
0026.txt
0027.txt
0028.txt
0029.txt
0030.txt
0031.txt
0032.txt
0033.txt
Загрузить папку и подпапки
пуск
Total C...
Micros...
диссер...
C++Bu...
text
EN
17:18

Подсистема хранения

Хранение данных системы,
таких как тексты и паспорта сообщений, лексический словарь, статистика запросов пользователей и прочая информация.

Подсистема выделения тем и вычисления актуальности

- Выделение тем
  - Подготовка кластерного анализа
  - Кластерный анализ абзацев
  - Переход от групп абзацев к группам сообщений
- Вычисление актуальности

Подсистема визуализации

- Получение данных
- Дополнительная обработка данных
- Выдача данных пользователю и реакция на запросы

Дата: 09/09/01 Число сообщений 2012

#	Дата темы	Число сообщ.	Описание темы	Близкая тема
1	07/09/01	26	В ХОДЕ ПРОВЕРКИ ОАО "МОСЭРГО" ВЫЯВЛЕНЫ НАРУШЕНИЯ В ЭНЕРГОСБЫТОВОЙ ДЕЯТЕЛЬНОСТИ ЭТОЙ КОМПАНИИ, СООБЩИЛ ПРЕДСТАВИТЕЛЬ РАО "ЕЭС РОССИИ"	23
2	07/09/01	34	Именно на основе этих принципов, указал Иванов, Россия оказывает поддержку руководству Македонии и готова взаимодействовать с международным сообществом. -0-	132
3	07/09/01	29	В ходе следствия Виктор Тихонов свою вину признал частично, говорится в обвинительном заключении.	95
4	07/09/01	36	В марте, наконец, суд приступил к слушаниям, но затем последовали новые перерывы и переносы заседаний из-за болезни "ведущего адвоката".	50
5	07/09/01	27	ЭНЕРГЕТИКАМ УДАЛОСЬ ПЕРЕЛОМИТЬ КРИЗИСНУЮ СИТУАЦИЮ В ОБЕСПЕЧЕНИИ ТОПЛИВОМ РЯДА РЕГИОНОВ ДАЛЬНЕГО ВОСТОКА	
6	06/09/01	43	Комитет муниципальных займов и развития фондовых рынков правительства Москвы направил в суд несколько исков о взыскании с "Медиа-Моста" в общей сложности более 6 млрд рублей в счет оплаты векселей. Векселя были выданы Москомзайму в счет долга "Мостбанка", на счетах которого находились средства комитета. Срок погашения векселей наступил весной этого года. Однако, когда Москомзайм в июне предъявил векселя к погашению, "Медиа-Мост" отказался их оплачивать. -0-	
7	06/09/01	62	ИГОРЬ ИВАНОВ ЗАЯВЛЯЕТ, ЧТО ПРАГМАТИЧЕСКИЙ КУРС ВНЕШНЕЙ ПОЛИТИКИ РОССИИ В ОТНОШЕНИИ СНГ НАЧИНАЕТ ПРИНОСИТЬ РЕАЛЬНУЮ ОТДАЧУ	19

Описание темы:

В ходе следствия Виктор Тихонов свою вину признал частично, говорится в обвинительном заключении.

Имя сообщения и его описание	Дата	Близость к теме
НОВОСИБИРСКИЙ СУД ЗАСЛУШИВАЕТ ОБВИНИТЕЛЬНОЕ ЗАКЛЮЧЕНИЕ ПО ДЕЛУ ВИКТОРА ТИХОНОВА, ОДНОГО ИЗ ОБВИНЯЕМЫХ В ОРГАНИЗАЦИИ ПОКУШЕНИЯ НА КЕМЕРОВСКОГО ГУБЕРНАТОРА АМАНА ТУЛЕЕВА В ходе следствия Виктор Тихонов свою вину признал частично, говорится в обвинительном заключении.	06/09/01	0.00843112
В ПЯТНИЦУ В НОВОСИБИРСКОМ ОБЛАСТНОМ СУДЕ ПРОДОЛЖАТСЯ СЛУШАНИЯ ПО ДЕЛУ ВИКТОРА ТИХОНОВА В пятницу, как ожидается, суд начнет допрос Виктора Тихонова. Он, напомним, первым предстал перед судом из всех обвиняемых в организации покушения на кемеровского губернатора.	08/09/01	0.00940864
В СРЕДУ В НОВОСИБИРСКОМ ОБЛАСТНОМ СУДЕ НАЧИНАЕТСЯ ПРОЦЕСС ПО ДЕЛУ ВИКТОРА ТИХОНОВА		