ОРГАНИЗАЦИЯ ЦЕЛЕНАПРАВЛЕННОГО АКТИВНОГО ПОИСКА НА ОСНОВЕ ОЦЕНКИ СИНТАГМАТИЧЕСКИХ И ПАРАДИГМАТИЧЕСКИХ АССОЦИАЦИЙ ТЕКСТОВЫХ СООБЩЕНИЙ

THE ORGANIZATION OF PURPOSEFUL ACTIVE SEARCH ON THE BASIS OF ASSESSMENT OF SYNTAGMATIC AND PARADIGMATIC ASSOCIATIONS OF TEXT MESSAGES

D. Akimov
A. Dyatchenkova
V. Sachkov

Summary. The work deals with methods and technologies used to actively search for information and analyze text messages in a natural language, taking into account the specifics of Internet communications for conscious dialogue and the means of presenting information in the profile of social networks for implementing a model for collecting target information using associative dictionaries.

Keywords: active search, context analysis, text messages, associative thesaurus, syntagmatic associations, paradigmatic associations, associative dictionaries, semantic core.

Акимов Дмитрий Александрович
К.т.н., МГТУ МИРЭА (Москва)
akimovdmitri@gmail.com
Дятченкова Анастасия Юрьевна
Аспирант, МГТУ МИРЭА (Москва)
futurama_07@bk.ru
Сачков Валерий Евгеньевич
Аспирант, МГТУ МИРЭА (Москва)
megawatto@gmail.ru

Аннотация. В работе рассматриваются методы и технологии, применяемые для активного поиска информации и анализа текстовых сообщений на естественном языке с учетом специфики Интернет коммуникаций для осознанного ведения диалога и средства представления информации в профиле социальных сетей для реализации модели сбора целевой информации с применением ассоциативных словарей.

Ключевые слова: активный поиск, контекстный анализ, текстовые сообщения, ассоциативный тезаурус, синтагматические ассоциации, парадигматические ассоциации, ассоциативные словари, семантическое ядро.

Введение

В настоящее время глобальная сеть интернет охватила все области человеческой жизни, с её развитием появляется все больше различных социальных сервисов, сетей и т.д., ключевой ценностью в которых выступает информация. В связи с чем возникает все больше потребности в поиске, сборе и управлении полезной информации для различных целей.

Представим, что имеет место следующая ситуация: посетители ресурсов обладают необходимой информацией, но не сообщают её в силу невостребованности со стороны остальных коммуникаторов ресурсов.

Предлагается программа, главным принципом которой является **активный поиск информации**. Под **активным поиском** понимается поиск, при котором поисковая программа, будучи зарегистрированной от имени человека-посетителя на ресурсе, позволяющим размещать комментарии, ведет диалог с посетителями по «интересующей её» проблеме, интегрирует ответы и отсылает результат Заказчику.

При этом копии подобных программ должны быть размещены на ресурсах, посетители которых географически распределены. В результате появляется возможность в режиме реального времени контролировать процессы (перемещение техники, людей, строительство сооружений) на заданных территориальных образованиях.

Образцы вопросов:

- Собираюсь с семьей завтра часам к 11 в город X, долго ли придется стоять на паромной переправе?
- Не подскажите, по дороге на Y сегодня проехать можно. Мне сказали, что её ремонтируют. Кто-нибудь из вас по ней проезжал в ближайшее время?

Вопросы с некоторой модификацией могут повторно задаваться при автоматическом выявлении новых посетителей ресурсов.

Для ведения диалога на естественном языке от лица какого-либо человека используется такой инструмент как виртуальная сущность (ВС). ВС — информационный аналог личности человека. Как «аватарка» (картинка, фотография или анимационное изображение) олицетворяет пользователя в сети, так и ВС создаётся по образу и подобию человека

МЕТОДИКА СИНТЕЗА ДИАЛОГА ДЛЯ ОБЩЕНИЯ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

По сути, в формальном описании, алгоритм общения ВС в сети можно свести к следующему:

1. Для каждой ВС задаётся список близких по тематике поиска чатов, форумов и блогов сети Интернет вместе с именами и паролями для входа.

2. ВС считывает странички из этого списка и анализирует интересы и специальные метрики (см. ниже).

3. Странички форумов и сайтов, как правило, выполнены на стандартном программном обеспечении и имеют почти стандартную структуру по тегам: имя посетителя, аватар посетителя, фраза (текст). ВС забирает эту страничку целиком и начинает применять себя к каждой имеющейся фразе — пытается ответить и самостоятельно оценить собственный ответ по заранее заданным критериям. При нахождении фраз, на которые, по его мнению, он способен дать более-менее приемлемый ответ, он забирает к себе и саму фразу и ответ на неё и включает их в свою таблицу диалоговых квантов. А в дальнейшем этот позаимствованный в ходе своего обучения ответ считает для себя более правильным, чем тот, что подготовил сам.

Диалоговый квант — это минимальная единица речевого общения, отражающая позиции всех участников диалога, но не более одного высказывания каждого участника.

В общем виде, текстовое содержимое сообщений может быть представлено в следующем виде:

$$T = \{\mathrm{w}_1, \mathrm{w}_2, ..., \mathrm{w}_i, ..., \mathrm{w}_n\},$$

где T — текстовое содержимое, w_i — слово, занимающее i-ую позицию всообщении.

Содержимое диалога может быть представлено в следующем виде:

$$T_{\mathrm{i}}^{\inf} = \{\mathrm{H}_i, \mathrm{L}_i, \mathrm{B}_i\},$$

где $T_{\mathrm{i}}^{\inf}$ — текст i-ого сообщения, H_i — заголовок i-ого диалога, L_i — лид i-го диалога, B_i — основное содержимое диалога.

Текст сообщения иногда сопровождаемыми лидами:

$$T^{nav} = \bigcup_{i=1}^{k} H_i[\mathrm{L}_i],$$

где T^{nav} — текст сообщения, H_i — заголовок i-го диалога, $[\mathrm{L}_i]$ — лид i-ого диалога, который может присутствовать или нет, в зависимости от конкретного ресурса, k — число тем диалога.

Учитывая формулы, текст диалога может быть представлен в виде объединения текстов сообщений, связанных с ней, за исключением основного содержимого:

$$T^{nav} = \bigcup_{i=1}^{k} (T^{inf} \setminus T^{body}),$$

где T^{body} — часть текста, удаляемая из исходного текста информационной страницы для формирования его представления.

Наличие в тексте описанных признаков не позволяет однозначно определить, является ли диалог связанным: они указывают на наличие определённых связей между сообщениями, которые могут не сохраняться на всём протяжении диалога. Однако, частое использование в диалоге средств связки значительно увеличивает вероятность того, что диалог является связанным, поэтому в дальнейшем мы будем называть описанные признаки формальными признаками связанности диалога. В рамках данной работы рассматриваются диалоги, состоящие из предложений, соответствующих принятым в языке грамматическим нормам, которые могут быть прочитаны и понятны пользователем. Для таких диалогов наблюдается следующая зависимость: чем реже в диалоге используются различные средства связки, тем менее связаны его части, т.е. диалог в целом является менее связанным.

В рамках решаемой задачи связанность диалога будет рассматриваться нами как наличие у него определённого набора формальных признаков:

$$E_T = \{\mathrm{s}_1, ..., \mathrm{s}_m\},\ s_i \in S_E,$$

где E_T — связанность диалога T, S_i — i-ый формальный признак связанности, S_E — множество формальных признаков связанности, m — число учитываемых формальных признаков связанности.

Введём двухместную операцию сравнения двух ветвей диалогов (контекстов), для определения более связанного из них:

$$F_{comp}(T_1, T_2) = F_{comp}(E_{T_1}, E_{T_2}) = \sum_{i=1}^{m} comp(As_i^{T_1}, As_i^{T_2});$$

$$comp(As_i^{T_1}, As_i^{T_2}) = \begin{cases} 1, As^{T_1} > As^{T_2} \\ 0, As^{T_1} = As^{T_2} \\ -1, As^{T_1} < As^{T_2} \end{cases};\quad As^T = \frac{Ns^T}{n^T},$$

где $F_{comp}(T_1, T_2)$ — операция сравнения степени связанности ветвей диалога T_1 и T_2, E_{T_j} — степень связанности диалога T_j, As_i^{Tj} — отношение числа вхождений i-ого формального признака связанности (Ns^T) в диалог T_j к общему числу слов в нём (n^T), $comp(As^{T_1}{}_i, As^{T_2}{}_i)$ — операция сравнения частоты вхождения i-ого формального признака связанности в диалогах T_1 и T_2 соответственно.

Результатом операции сравнения $F_{comp}(T_1, T_2)$ является число — сумма результатов поэлементного сравнения соответствующих вхождений формальных признаков в диалогах T_1 и T_2. Полученный результат может быть интерпретирован следующим образом:

$$\begin{cases} E_{T_1} > E_{T_2}, & \mathrm{F}_{comp}(T_1, T_2) > 0 \\ E_{T_1} \approx E_{T_2}, & \mathrm{F}_{comp}(T_1, T_2) = 0, \\ E_{T_1} < E_{T_2}, & \mathrm{F}_{comp}(T_1, T_2) < 0 \end{cases}$$

При сравнении диалогов с учётом вышеизложенных формул операция сравнения примет вид:

$$F_{comp}(T^{\inf}, T^{nav}) =$$

$$= F_{comp}(\{\mathrm{H}^{\inf}, \mathrm{L}^{\inf}, \mathrm{B}^{\inf}\}, \sum_{i=1}^{n} \{\mathrm{H}_i^{\inf}, \mathrm{L}_i^{\inf}\}),$$

где $\mathrm{H}_i^{\inf}$ — заголовок i-ого диалога, а $\mathrm{L}_i^{\inf}$ — лид i-ого диалога, используемый в диалоге T^{nav}.

Введённая операция сравнения позволяет определить более связанный диалог из двух диалогов, однако для решения поставленной задачи классификации должна быть введена числовая мера связанности диалога. Ранее мы определили, что чем чаще в диалоге встречаются средства связи, тем больше вероятность того, что диалог в целом является связанным — используем данный подход для количественного выражения степени связанности диалога:

$$E_T = \sum_{i=1}^{m} Ns_i^T; \quad E_{T\inf} > E_{Tnav},$$

где E_T — числовое значение степени связанности диалога.

Выделенные ранее формальные признаки связанности диалога могут быть разделены на две группы по способу их определения: на признаки, выражаемые при помощи определённых частей речи и признаки, выражаемые при помощи использования схожих словоформ.

Признаками первой группы является использование в диалоге деепричастий, местоименных существительных, наречий, числительных и союзов. Числовым представлением частоты использования является отношение числа появления соответствующей части речи к общему числу слов в диалоге:

$$p_i^s = \frac{Ns_i}{n}; \quad i = \{1, ..., \mathrm{k}\},$$

где p_i^s — числовое представление s_i-ого признака связанности, определяемого по количеству использований частей речи, Ns_i — число использований в диалоге частей речи, соответствующих s_i-ому признаку, n — общее число слов в диалоге, k — число признаков связанности, относящихся к первой группе.

Признаками второй группы является использование словоформ имеющих одинаковый корень, а также разных словоформ с одним корнем. Числовое представление этих формальных признаков может быть получено через отношения числа повторяющихся словоформ к общему числу слов:

$$r_j^s = \frac{Nwd}{n}, \quad Nwd = (\mathrm{n} - n_u), \quad \mathrm{j} = \{1, 2\},$$

где r_j^s — числовое представление s_j-ого признака связанности, определяемого по количеству повторяющихся словоформ, Nwd — число повторяющихся словоформ, соответствующих s_j-ому признаку связанности, n — общее число слов в диалоге, n_u — число уникальных слов в диалоге.

Учитывая формулы, для определения степени связанности диалога, исходный диалог страницы должен быть представлен в следующем виде:

$$T = \{\mathrm{p}_1^s, ..., \mathrm{p}_k^s, \mathrm{r}_1^s, \mathrm{r}_2^s\} = [\mathrm{p}_1^s, ..., \mathrm{p}_k^s, \mathrm{r}_1^s, \mathrm{r}_2^s]$$

Таким образом, предлагаемая модель диалогового взаимодействия учитывает 8 признаков связанности, выявляемых на морфологическом уровне анализа: частоту использования деепричастий, местоимённых существительных, местоименных прилагательных, общего числа местоимений, числительных, союзов, а также полных дейктических и лексико-семантических повторов.

Ассоциативный тезаурус

В задаче построения связанного диалога важной проблемой является понимание смысла сообщений. В социальных сетях, в основном, используются короткие сообщения, по этой причине выделение семантической информации из сообщения представляется еще более затруднительно чем при семантической разметки больших таксовых документов. Проблема усугубляется применением жаргонных слов и словосочетаний. Для решения проблемы семантической разметки сообще-

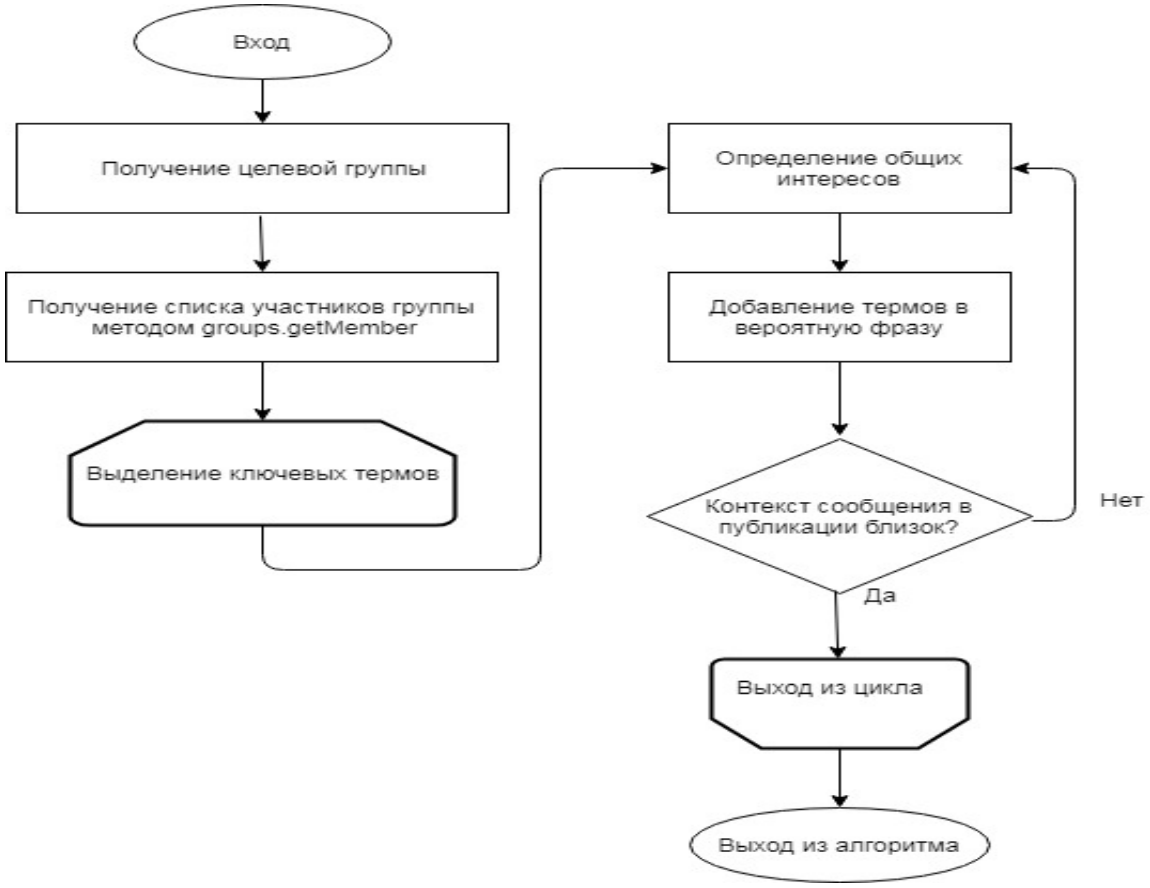

Рис. 1. Алгоритм трансляции заданной темы в диалоговую последовательность

ний и выделения смысловых конструкций предлагается использовать ассоциативные словари. Словесные ассоциации, используемые в ассоциативных словарях, основаны на подсознательной способности человека выстраивать логические связи на основании знаний об окружающем мире. Так например слово «море» может ассоциироваться с «чайками» и «кораблями», при этом если в сообщении встречается описание событий со словами «чайки» и «корабль» можно предположить что описываемые события происходят в море.

Ассоциативный словарь — словарь в котором отражена информация о том, как говорящие соединяют слова-реакции с определенными словами-стимулами, что отражает семантические парадигматические и синтагматические (линейные) связи.

Ассоциативный эксперимент — это прием, направленный на выявление ассоциаций, сложившихся у индивида в его предшествующем опыте.

Для создания ассоциативного словаря используется «Ассоциативный эксперимент», который является наиболее разработанной техникой психолингвистического анализа семантики. Процедура ассоциативного эксперимента следующая: испытуемым предъявляется список слов и говорится, что им необходимо ответить первыми приходящими в голову словами. В среднем, каждому испытуемому даётся 100 слов и 7–10 минут на ответы.

Существует несколько разновидностей ассоциативного эксперимента:

- Свободный ассоциативный эксперимент. Испытуемым не ставится никаких ограничений на реакции.
- Направленный ассоциативный эксперимент. Испытуемому предлагается давать ассоциации определённого грамматического или семантического класса (например, подобрать прилагательное к существительному).
- Цепочечный ассоциативный эксперимент. Испытуемым предлагается реагировать на стимул несколькими ассоциациями — например, дать в течение 20 секунд 10 реакций.

При анализе ответов ассоциативного эксперимента выделяют, прежде всего синтагматические и парадигматические ассоциации:

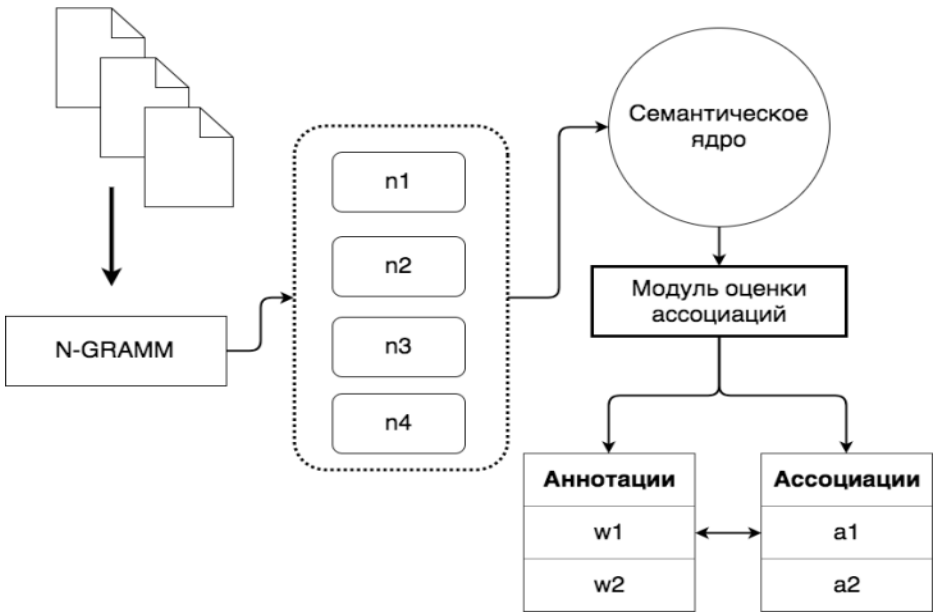

Рис. 2. Структура логического вывода алгоритма

- небо — голубое, машина — едет, курить — плохо
- стол — стул, отец — мать

Синтагматическими ассоциациями называются ассоциации, грамматический класс которых отличен от грамматического класса слова-стимула. Парадигматические ассоциации представляют собой слова-реакции того же грамматического класса, что и слова-стимулы.

Они подчиняются принципу «минимального контраста», согласно которому чем меньше отличаются слова-стимулы от слов-реакций по составу семантических компонентов, тем более высока вероятность актуализации слова-реакции в ассоциативном процессе. Этот принцип объясняет, по чему, по характеру ассоциаций можно восстановить семантический состав слова-стимула: множество ассоциаций, выданных на слово, содержит ряд признаков, аналогичных содержащимся в слове-стимуле.

Носитель языка по реакциям может достаточно легко восстанавливать стимул, например, «каникулы»:

Летние, лето, отдых, короткие, скоро, ура, безделье, в Простоквашино, начались, школа.

Ассоциативный эксперимент даёт возможность построить семантическую структуру слова. Он служит ценным материалом для изучения психологических эквивалентов того, что в лингвистике называется семантическим полем, и вскрывает объективно существующие в психике носителя языка семантические связи слов.

Минус данного подхода в том, что ассоциативный словарь составляется вручную. Занимает много времени и имеет высокие трудозатраты. Имеет огромное значение автоматизировать данный процесс.

Алгоритм трансляции в диалоговую последовательность

Цель алгоритма трансляции состоит в том, чтобы разложить сообщение и извлечь всю соответствующую информацию о структуре и контенте, чтобы обеспечить выделение ключевых смыслов. Включая интерпретацию в соответствии с контекстами, языками или читателями.

Например, слово «багажник» может относиться к области хранения (в контексте транспортных средств), к коробке для хранения одежды (в контексте путешествия) или к слону (в контексте сафари), пример рисунок 1. Стрелки представляют собой параметры, связанные с отношениями. Для связанных слов может быть несколько значений, и только кластеризация слов обеспечивает важный контекст, который предоставляет читателям смысл; Например, Safari также является именем интернет-браузера.

Для нашего алгоритма мы создадим семантическое ядро, которое будет, основываться на корпусе описательных документов и позволит выполнять семантическую разметку сообщений, относящиеся к данной тематике.

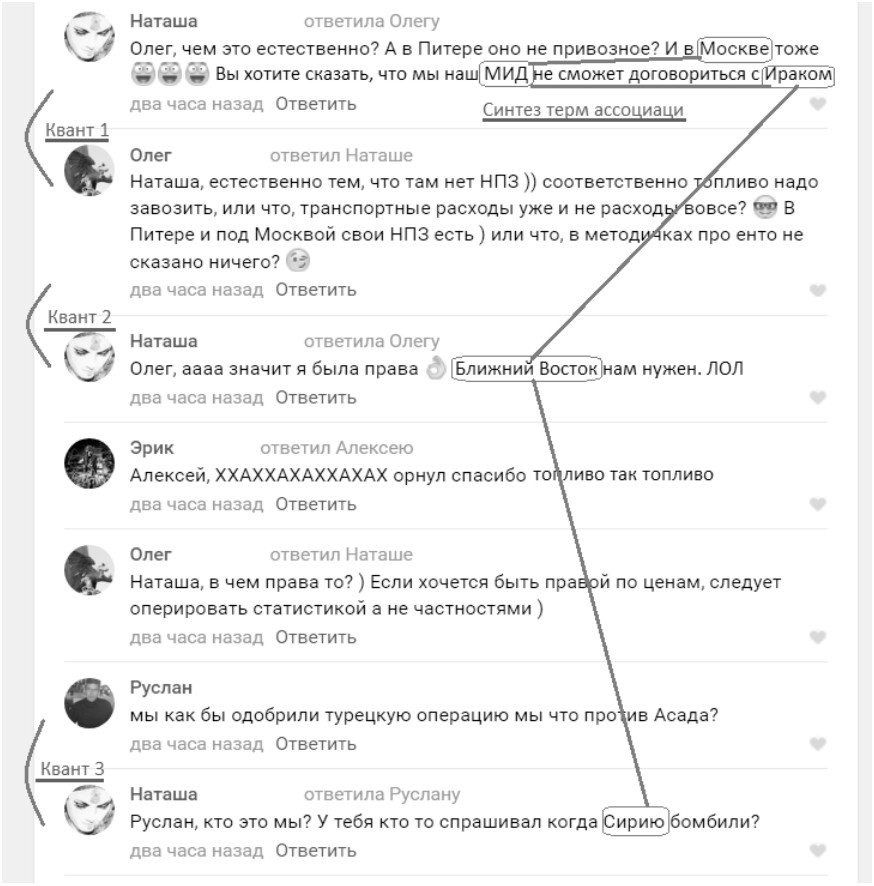

Рис. 3. Результат формирования сообщений с трансляцией тезисов и словосочетаний интересующей новости в диалоговую последовательность

```
Bot  >>> Хорошо  Watto  поговим о новостях, что вы хотите узнать?
User >>> расскажи мне о крушении самолета в донбассе на украине
Bot  >>> Хорошо  Watto  ищу:  расскажи мне о крушении самолета в донбассе на украине
Bot  >>> Я думаю мы говорим о политике?
User >>> да
Bot  >>> Хорошо, посмотрим что я смогу найти...
Bot  >>> Смотрю документ 19996 из 19997
                                          doc_num          EMD
                                                    text
18348    18348   5.114148  москв 18 – причин крушен малайзийск лайнер нес...
18331    18331   5.114148  москв 18 – причин крушен малайзийск лайнер нес...
18365    18365   5.114148  москв 18 – причин крушен малайзийск лайнер нес...
18343    18343   5.360942  москв 18 – заявлен киевск власт крушен малайзи...
18315    18315   5.360942  москв 18 – заявлен киевск власт крушен малайзи...
18326    18326   5.360942  москв 18 – заявлен киевск власт крушен малайзи...
18360    18360   5.360942  москв 18 – заявлен киевск власт крушен малайзи...
18333    18333   5.687893  москв 18 – лидер справедлив росс серг мирон уб...
18350    18350   5.687893  москв 18 – лидер справедлив росс серг мирон уб...
18367    18367   5.687893  москв 18 – лидер справедлив росс серг мирон уб...
18164    18164   6.503808  москв 28 – основан полага ход расшифровк черн ...
18366    18366   6.677729  москв 18 – росс готов оказа логистическ содейс...
18332    18332   6.677729  москв 18 – росс готов оказа логистическ содейс...
18349    18349   6.677729  москв 18 – росс готов оказа логистическ содейс...
18215    18215   6.754733  москв 24 – сша предостав документирова дан под...
```

Рис. 4. Диалог между пользователем и системой

Примечание: *n*-граммы.— другая модель для упрощения распознавания содержание текста; n-граммная модель определяет и сохраняет смежные последовательности слов в тексте;

Конечным результатом отработки алгоритма является диалог между системой и пользователем, в ходе которого подобраны и определены документы по затрагиваемой тематике.

Вывод

Создан прототип диалоговой поисковой системы способной анализировать публикации новостей на уровне семантики. Система задействует 8 признаков связности, на основании которых определяется диалог с пользователем, в ходе которого извлекается информация и смысл разговора. После чего, система определяет подходящие новостные документы.

ЛИТЕРАТУРА

1. Matt Kusner, Yu Sun, Nicholas Kolkin, Kilian Weinberger From Word Embeddings To Document Distances // Proceedings of the 32nd International Conference on Machine Learning, PMLR37:957–966, 2015.
2. Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. // IEEE International Conference on Computer Vision, pages 59–66, January 1998.
3. Van Dijk T. A. Critical Discourse Analysis // Handbook of Discourse Analysis / D. Tannen, D. Schiffrin, H. Hamilton (eds). Oxford: Blackwell, 2001
4. Акимов Д. А., Сачков В. Е., Алёшкин А. С., Уманский В. И. Обработка и компьютерный анализ информации об опубликованных уязвимостях нулевого дня на естественных языках. Проблемы информационной безопасности. Компьютерные системы, 2017, № 2, с. 9–15.
5. А.С. Сигов, Д. А. Акимов, Д. О. Жуков, Е. Г. Андрианова, В. Е. Сачков, В. К. Раев Психолингвистический анализ русскоязычных текстовых сообщений на основе их фоносемантических статистических характеристик. Информатика и её применения, 2017. Т. 11, Вып. 3. с. 77–86
6. D. Akimov, P. Krug, A. Ostroukh, E. Matiukhina, V. Ivchenko «Development of an automobile robot system model based on soft computing in an unsteady environment» ARPN Journal of Engineering and Applied Sciences VOL. 12, NO. 11, JUNE2017
7. Д.А. Акимов, Д. А. Потапов, «Выявление речевых конструкций повышающих точность работы системы информационного поиска» Современная наука: актуальные проблемы теории и практики. 2017 № 1 с 41–43
8. Д.А. Акимов, В. В. Котельников, Д. А. Скоселева, А. Ю. Дятченкова, «Прогнозирование остаточного ресурса на основе мягких вычислений» Современная наука: актуальные проблемы теории и практики. 2017 № 1 с 20–22

© Акимов Дмитрий Александрович (akimovdmitri@gmail.com),
Дятченкова Анастасия Юрьевна (futurama_07@bk.ru), Сачков Валерий Евгеньевич (megawatto@gmail.ru).
Журнал «Современная наука: актуальные проблемы теории и практики»

МГТУ МИРЭА