УДК 004.04:81

ТЕМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ СТАТЕЙ НОВОСТНОГО РЕСУРСА МЕТОДАМИ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА

Толмачев Р.В., Воронова Л.И.

Ордена Трудового Красного Знамени ФГБОУ ВО «Московский технический университет связи и информатики», Москва, e-mail: voronova.lilia@ya.ru

Решение задачи тематической классификации текстов в сети Интернет включает в себя ряд последовательных этапов: извлечение данных из ресурса в сети Интернет, их обработку математическими методами и наглядное представление результата. В настоящее время автоматическая классификация становится все более важной задачей, как с теоретической, так и с прикладной точек зрения, и имеет широкий диапазон применения. Одно из направлений – тематическая классификация, предполагает автоматическое определение темы новостной статьи, реализованное с помощью методов семантического анализа текста. Целью работы является исследование и реализация методов латентно-семантического и вероятностного латентно-семантического анализа, для определения тематики новостных статей. В работе описана реализация указанных методов и сравнительная проверка их работоспособности на нескольких примерах.

Ключевые слова: тематическое моделирование, латентно-семантический анализ, вероятностный латентно-семантический анализ, новостные статьи

THEME CLASSIFICATION OF NEWS ARTICLES BY METHODS OF LATENT SEMANTIC ANALYSIS

Tolmachev R.V., Voronova L.I.

Moscow Technical University of Communication and Informatics, Moscow, e-mail: voronova.lilia@yandex.ru

The problem of thematic classification of the Internet texts include a number of tasks: the extraction of data from a resource on the Internet, processing this data by mathematical methods and presentation of results. In addition, automatic classification is becoming more important now days. Automatic theme identification of news article can be implemented using text semantic analysis methods. The aim of this work is to research and implement two methods (latent semantic method and probabilistic latent semantic method to analysis) to determine the subject of news articles. The results of this work show that these methods can be used for thematic classification of news articles.

Keywords: topic model, latent semantic analysis, probabilistic latent semantic analysis, news articles

В наше время интернет развивается в быстром темпе. Растет количество информации в сети и количество пользователей. В 2015 г. российская ежемесячная интернет-аудитория выросла на 9,2% до 80,5 млн пользователей, говорится в докладе Российской ассоциации электронных коммуникаций [9].

Неструктурированные данные составляют большую часть информации, с которой имеют дело пользователи. Поэтому автоматическая кластеризация (выявление похожих по темам текстов) является одной из важнейших задач, решаемых с помощью информационных систем [1].

На рынке программного обеспечения представлено большое количество программ, реализующих классификацию текстов. Российская служба Яндекс.Новости автоматически группирует данные в новостные сюжеты и составляет аннотации этих новостных статей [4].

Для автоматического определения темы используются методы тематического моделирования, которые реализуются с помощью различных алгоритмов.

В статье описано применение методов тематического моделирования для автоматического определения темы новостных статей, взятых с сайта «РИА Новости». Рассмотрены: принцип сбора данных через веб-интерфейс новостного сайта, задача построения тематической модели, два метода латентно-семантического анализа, проведена их алгоритмизация и программная реализация.

Методика сбора данных из новостных сайтов

Веб-интерфейсы новостных статей являются источниками данных реального времени. Для взаимодействия с ними можно использовать *W*eb браузер или специальные приложения. Из-за того что новостные сайты не предполагают возможность сбора данных, то возникает проблема – разная структура данных для разных новостных источников. С другой стороны, это никак не влияет на структуру новостного текста [6].

Изучив ряд новостных источников в сети Интернет, таких как «РИА Новости», РБК, KP.RU и других (по рейтингу их популярности [8], можно представить структуру

новостной записи [10]: заголовок – отражает тему новостной статьи, состоит из небольшого набора слов, и основные факты – содержат главную тему, состоит из первых 1 или 2 абзацев текста новости. Оставшийся текст новости, как правило, содержит детали, которые только косвенно связаны с темой новости.

Для извлечения текста новости с вебстраницы новостного сайта существует множество подходов. Ниже представлены три из них, как имеющие большое практическое применение в современных технологиях:

1. Извлечение данных с использованием только html-кода документа. Этот процесс включает три этапа [5].

– Получение исходного кода вебстраницы. В разных языках для этого предусмотрены различные способы. Например, в PHP чаще всего используют библиотеку с URL или же встроенную функцию file_get_contents.

– Извлечение из html-кода необходимых данных. Получив страницу, необходимо обработать её – отделить обычный текст от гипертекстовой разметки, для этого можно использовать регулярные выражения, а также специализированные библиотеки [12].

– Фиксация результата. Полученные данные сохраняются в базе данных или в отдельных текстовых файлах.

2. Извлечение данных с использованием алгоритмов компьютерного зрения. Это точный алгоритм, однако самый сложный и ресурсоемкий. Он включает два этапа:

– Рендеринг страницы. Рендеринг (англ. rendering – «визуализация») – это термин в компьютерной графике, обозначающий процесс получения изображения по модели с помощью компьютерной программы. На этом этапе алгоритм преобразует модель, web-страницу, в изображение.

– Извлечение данных. С помощью алгоритмов компьютерного зрения (англ. Computer Vision) происходит детектирование контента интернет-страницы: текста, картинок и их расположения на странице.

3. Извлечение данных на уровне сайта целиком. С помощью формул и таблиц Google или с помощью дополнительных приложений можно найти на странице необходимый участок кода, в котором заключена необходимая информация. Эти части кода повторяются в пределах одной страницы и на других страницах, имеющих аналогичную структуру. Данный подход помогает выделить и импортировать повторяющиеся данные автоматически, существенно сэкономив время и предупредив возможные ошибки копирования этой информации вручную [7].

Задача построения тематической модели

Тематическая модель – это модель, которая определяет принадлежность документа, из некоторого набора документов к определенной теме.

С математической точки зрения задача классификации текстов по темам сводится к задаче одновременной классификации набора документов по одному и тому же множеству тем [3].

При построении тематической модели, вводится ряд обозначений и предположений.

Пусть D это набор документов, содержащий множество документов d

$$D = \{d_1, d_2, d_3, ..., d_n\},$$

где n – это количество документов в наборе.

Каждый документ d в свою очередь представляет собой множество слов W_d.

$$W_d = \{w_1, w_2, w_3, ..., w_{d_n}\},$$

где $|W_d|$ – это количество слов в документе.

Каждый документ d имеет тему t, принадлежащую множеству тем T

$$T = \{t_1, t_2, t_3, ..., t_n\}.$$

Предполагается, что существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая неизвестна. Коллекция документов рассматривается как множество троек (d, w, t), выбранных случайно и независимо из дискретного распределения p (d, w, t), заданного на конечном множестве $D \times W \times T$. Документы $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, тема $t \in T$ является латентной (скрытой) переменной [2]. Кроме этого:

– Порядок документов в наборе D не имеет значения.

– Порядок слов в документе d не имеет значения.

– Слова, встречающиеся часто в большинстве документов, не важны для определения тематики.

– Каждая тема $t \in T$ описывается неизвестным p ($w \mid t$) на множестве слов $w \in W$.

– Каждый документ $d \in D$ описывается неизвестным распределением p ($t \mid d$) на множестве тем $t \in T$.

Построить тематическую модель – значит, найти матрицы

$$\Phi = \|p(w|t)\| \Phi = \|p(w|t)\|$$

и

$$\Theta = \|p(t|d)\| \Theta = \|p(t|d)\|$$

по коллекции D.

Методы латентно-семантического анализа

Метод латентно-семантического анализа LSA (англ. Latent semantic analysis, LSA)

Метод LSA был запатентован в 1988 г. группой американских инженеров-исследователей. Впервые метод был применен для автоматического индексирования текстов, выявления семантической структуры текста. Затем этот метод был довольно успешно использован для представления баз знаний и построения когнитивных моделей. В США этот метод был запатентован для проверки знаний школьников и студентов, а также проверки качества обучающих методик [13]. Метод основан на сингулярном разложении матриц SVD (англ. Singular value decomposition, SVD).

Этапы проведения метода LSA:

● Составляется общий словарь всех уникальных слов во всех документах, без учета слов не несущих смысловой нагрузки (стоп-слов).

● Для каждого слова определяется частота его вхождения в каждый документ v_{ij}.

● Составляется терм-документная матрица **A**. В терм-документной матрице строки соответствуют документам в коллекции, а столбцы соответствуют терминам.

Таблица 1

Представление терм-документной матрицы

	w_1	...	w_m
d_1	$v_{1,1}$		$v_{1,m}$
...			
d_n			$v_{n,m}$

Производится сингулярное разложение матрицы $A = USV^T$.

Полученные в результате матрицы U и V являются искомыми матрицами

$$\Phi = \left\| p\left(w|t\right) \right\| \Phi = \left\| p\left(w|t\right) \right\|$$

и

$$\Theta = \left\| p\left(t|d\right) \right\| \Theta = \left\| p\left(t|d\right) \right\|$$

соответственно. На рис. 1 представлен алгоритм метода LSA.

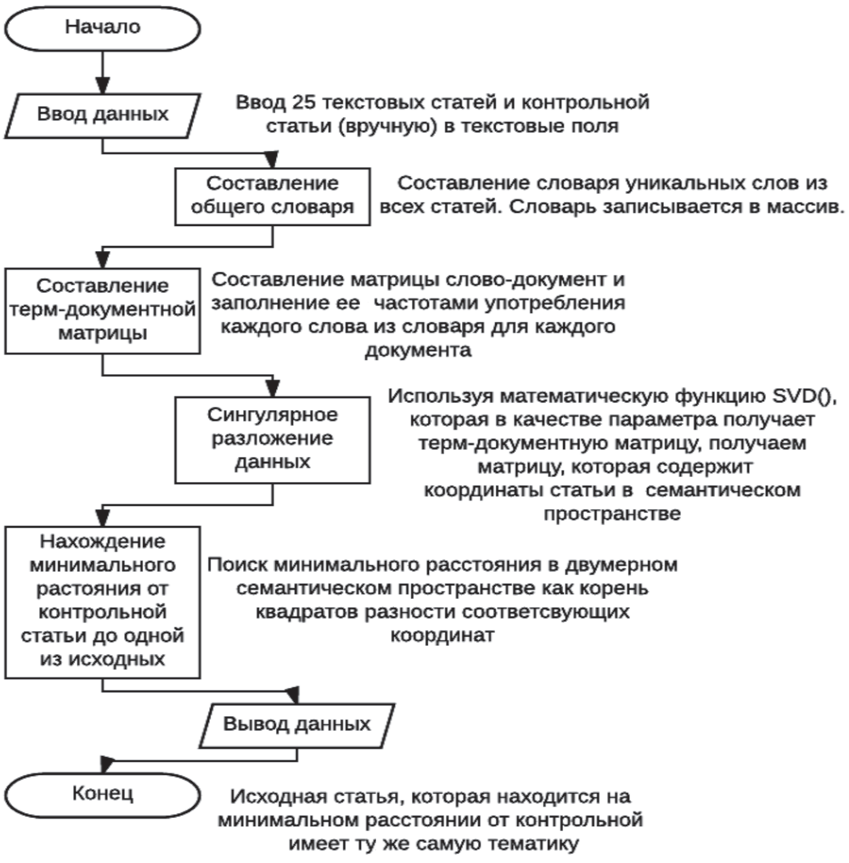

Рис. 1. Алгоритм LSA метода

Метод вероятностного латентно-семантического анализа PLSA (англ. Probabilistic latent semantic analysis, PLSA)

Данный метод является дальнейшим развитием латентно-семантического анализа. ВЛСА применяется в таких областях, как информационный поиск, обработка естественного языка, машинное обучение, и смежных областях. Данный метод был впервые опубликован в 1999 г. Thomas Hofmann [11]. Метод основан на применении EM-алгоритма (алгоритм поиска оценок максимального правдоподобия).

Рассмотрим вероятностную тематическую модель

$$p(D,\Phi,\Theta),$$

где $\Phi=(\varphi_{wt})_{W\times T}$ – искомая матрица терминов тем, $\varphi_{wt}\equiv p(w|t)$, $\Theta=(\theta_{td})_{T\times D}$ – искомая матрица тем документов, $\theta_{td}\equiv p(t|d)$.

Для вычисления значений φ_{wt} и θ_{td} используется двухшаговый EM-алгоритм.

E-шаг. На этом шаге, используя текущие значения параметров φ_{wt} и θ_{td}, по формуле Байеса вычисляется значение условных вероятностей $H_{dwt}\equiv p(t|d,w)$ для всех тем $t\in T$ для каждого термина $w\in d$ для всех документов $d\in D$:

$$H_{dwt}\equiv p(t\,|\,d,w)=\frac{p(w\,|\,t)p(t\,|\,d)}{p(w\,|\,d)}=\frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s}\varphi_{ws}\theta_{sd}}.$$

M-шаг. На этом шаге решается обратная задача: по условным вероятностям тем H_{dwt} вычисляются новые приближения φ_{wt} и θ_{td}.

Величина $\hat{n}_{dwt}=\hat{n}_{wt}p(t|d,w)=n_{dzw}H_{dwt}$ оценивает число n_{dwt} вхождений термина w в документ d, связанных с темой t. При этом оценка не всегда является целым числом. Просуммировав $\hat{n}_{dwt}$ по документам d и по терминам w, получим оценки:

$$\hat{n}_{wt}=\sum_{d\in D}\hat{n}_{dwt},\hat{n}_{t}=\sum_{w\in W}\hat{n}_{wt},$$

$$\hat{n}_{dt}=\sum_{w\in W}\hat{n}_{dwt},\hat{n}_{d}=\sum_{t\in T}\hat{n}_{dt}.$$

В соответствии с формулами, вычисляются частотные оценки условных вероятностей φ_{wt} и θ_{td}:

$$\varphi_{wt}=\frac{\hat{n}_{wt}}{\hat{n}_{t}}\ \text{и}\ \theta_{td}=\frac{\hat{n}_{dt}}{\hat{n}_{d}}.$$

Псевдокод PLSA изображен на рис. 2 [3]:

Вход: коллекция документов D, число тем $|T|$;
Выход: Φ, Θ;
инициализировать вектор-столбцы ϕ_t, θ_d случайным образом;
повторять
  обнулить n_{wt}, n_{td}, n_t, n_d для всех $d\in D,\ w\in W,\ t\in T$;
  для всех $d\in D,\ w\in d$
    $p(w\,|\,d):=\sum_{t\in T}\phi_{wt}\theta_{td}$;
    для всех $t\in T$
      $p(t\,|\,d,w):=\phi_{wt}\theta_{td}/p(w\,|\,d)$;
      увеличить n_{wt}, n_{td}, n_t, n_d на $n_{dw}p(t\,|\,d,w)$;
  $\phi_{wt}:=n_{wt}/n_t$ для всех $w\in W,\ t\in T$;
  $\theta_{td}:=n_{td}/n_d$ для всех $d\in D,\ t\in T$;
пока Θ и Φ не сойдутся;

Рис. 2. Псевдокод PLSA метода

Таблица 2

Результаты работы программы

Раздел сайта	Http-адрес новости	LSA	PLSA
Политика	https://ria.ru/politics/20161110/1481118723.html	Политика	Политика
Экономика	https://ria.ru/economy/20161110/1481125019.html	Политика	Экономика
Спорт	https://ria.ru/sport/20161128/1482282838.html	Спорт	Спорт
Наука	https://ria.ru/science/20161128/1482296195.html	Политика	Наука
Культура	https://ria.ru/culture/20161124/1482106648.html	Культура	Культура

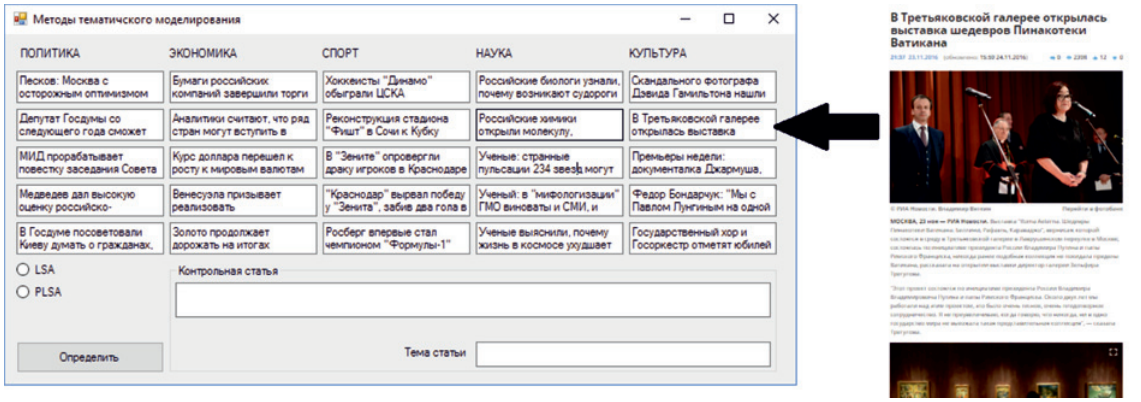

Рис. 3. Графический интерфейс программы

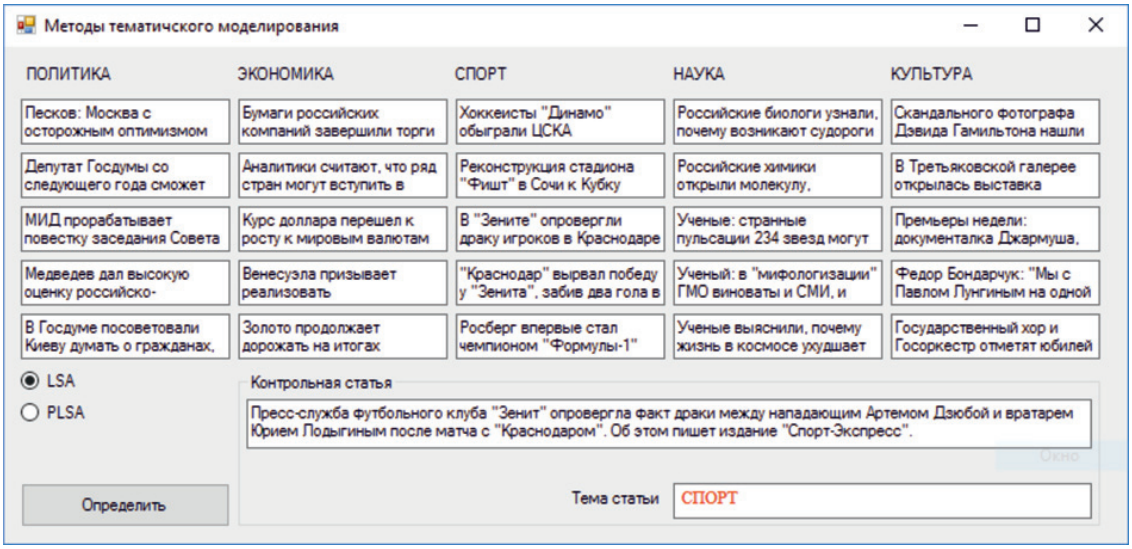

Рис. 4. Результаты работы программы LSA

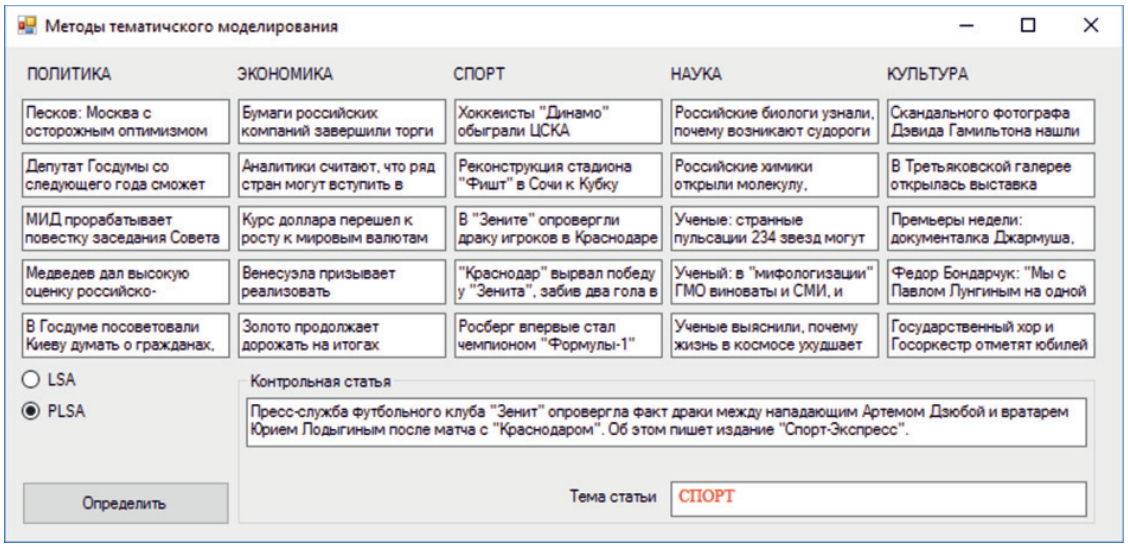

Рис. 5. Результаты работы программы PLSA

Программная реализация

Для реализации методов была разработана программа на языке программирования C#, ее графический интерфейс изображен на рис. 3.

Программа получает на вход 25 текстов. Текстами являются новостные статьи, взятые из новостного источника в сети Интернет. На данном сайте тексты отсортированы по темам (разделам сайта). Тексты взяты из разных разделов (по 5 текстов на тему): Политика, Экономика, Спорт, Культура, Наука.

Тексты новостей извлекаются с использованием обработки html-кода веб-страницы новостного сайта. Задачей программы является определение тематики контрольной статьи.

Для проверки работоспособности программы были выбраны статьи с предопределенной темой. Из каждого раздела по одной. В табл. 2 представлены результаты работы программы для этих статей.

На рис. 4 и 5 представлены результаты работы программы, для статьи из раздела «Спорт», методами LSA и PLSA соответственно.

Выводы

Можно констатировать, что тематики контрольных статей отобразились верно, при использовании метода PLSA, что подтверждает возможность его использования для классификации тематик новостных статей. Метод LSA допускает ошибки в определении темы, в связи с тем, вероятно, что обучающий набор слишком мал. Автоматическая категоризация текстов весьма актуальна для поисковых систем и систем обработки и извлечения знаний.

Список литературы

1. Воронова Л.И., Исаков В.А., Катина Т.С. Проектирование информационной системы организации мониторинга и контроля качества образовательных программ для методического управления РГГУ // В сборнике: Современные информационные технологии в профессиональной деятельности СИТ – 2015 Московский финансово-юридический университет МФЮА. 2015. – С. 27–33.

2. Воронцов К.В. Вероятностные тематические модели коллекции текстовых документов. URL: www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (дата обращения: 10.10.2016).

3. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды ИСП РАН. 2012.

4. О сервисе Яндекс-Новости [Электронный ресурс]. – Режим доступа: yandex.ru/support/news/ (дата обращения: 10.11.2016).

5. Парсинг html-сайтов с помощью PHP, Ruby, Python [Электронный ресурс]. – Режим доступа: http://parsing.valemak.com/ru/what-why-how/stages-of-parsing/(дата обращения: 09.11.2016).

6. Пестряев А.А., Воронова Л.И., Воронов В.И. Проектирование мультиагентной системы для сбора текстовой информации из сети // Вестник РГГУ. Серия: Документоведение и архивоведение. Информатика. Защита информации и информационная безопасность. – 2015. – № 12 (155). – С. 43–56.

7. Пестряев А.А., Воронова Л.И. Мультиагентная система.взаимодействия агента-собирателя с базой данных// Современные наукоемкие технологии. – 2014. – № 5–2. – С. 214–217.

8. Рейтинг новостных порталов. [Электронный ресурс]. – Режим доступа: www.liveinternet.ru/rating/ru/media/ (дата обращения: 01.11.2016).

9. Российский интернет прибавил темпы роста [Электронный ресурс]. – Режим доступа: https://www.vedomosti.ru/technology/articles/ 2016/04/13/637574-rossiiskii-internet-tempi-rosta (дата обращения: 18.11.2016).

10. Солошенко А.Н., Орлова Ю.А., Розалиев В.Л. Автоматическое выделение сюжетов и тем из потока новостных сообщений // Известия ВолгГТУ. – 2015. – № 2 (157).

11. Thomas Hofmann, Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in en: Information Retrieval (SIGIR-99), 1999.

12. Voronova L.I. «EXPERT LIST» subsystem design for methodical department of RSUH // International journal of applied and fundamental research. – 2015. – № 2. – С. 24930.

13. Readings in Latent Semantic Analysis for Cognitive Science and Education [Электронный ресурс]. – Режим доступа: lpnc.univ-grenoble-alpes.fr/resources/benoit_lemaire/lsa.html (дата обращения: 10.11.2016).