УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2016 Том 16 № 2 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2016 Vol. 16 No 2 ISSN 2226-1494 http://ntv.ifmo.ru/en

**УДК 004.912**

# ОЦЕНКА СЕМАНТИЧЕСКОЙ БЛИЗОСТИ ПРЕДЛОЖЕНИЙ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ МЕТОДАМИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

**А.Е. Письмак[a], А.Е. Харитонова[a], Е.А. Цопа[a], С.В. Клименков[a]**

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: Evgenij.Tsopa@cs.ifmo.ru

**Информация о статье**
Поступила в редакцию 26.01.16, принята к печати 05.02.16
doi:10.17586/2226-1494-2016-16-2-324-330
Язык статьи – русский

**Ссылка для цитирования:** Письмак А.Е., Харитонова А.Е., Цопа Е.А., Клименков С.В. Оценка семантической близости предложений на естественном языке методами математической статистики // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 2. С. 324–330. doi:10.17586/2226-1494-2016-16-2-324-330

**Аннотация**

**Предмет исследования.** Рассмотрены особенности структурной организации статей открытого словаря Wiktionary в контексте его использования в качестве источника для построения семантической сети. Изучены рекомендации сообщества по оформлению статей, основные шаблоны и особенности оформления реальных словарных статей. Рассмотрена проблема численной оценки семантической близости структурных элементов статей Wiktionary. Проведен анализ существующих программных продуктов для определения семантической близости таких элементов, изучены алгоритмы их функционирования, определены их преимущества и недостатки. **Методы.** Использованы методы математической статистики, учитывающие некоторые специфичные для Wiktionary форматы представления данных. Предложен метод определения семантической близости на основании статистических данных сравниваемых структурных элементов. **Основные результаты.** Сделаны выводы о невозможности прямого использования статей Wiktionary в качестве основы для построения семантической сети и о необходимости выявления скрытых структурных связей, для чего было предложено использовать метод оценки семантической близости предложений. Получен алгоритм, позволяющий на основе набора исходных предложений вычислить коэффициенты достоверности того, что каждая пара предложений является семантически близкой. Исследование количественных и качественных характеристик разработанного алгоритма показало его существенное преимущество над существующими решениями в производительности при несколько меньшей точности оценки семантической близости. **Практическая значимость.** Полученный алгоритм может быть полезен при разработке инструментов автоматического разбора словаря Wiktionary, а также при решении задач определения семантической близости небольших фрагментов текста на естественном языке в случае, если требования к производительности алгоритма являются более критичными, чем требования к его точности.

**Ключевые слова**
семантическая близость, математическая статистика, множества, токены, Wiktionary, семантический анализ, текст

# EVALUATION OF SEMANTIC SIMILARITY FOR SENTENCES IN NATURAL LANGUAGE BY MATHEMATICAL STATISTICS METHODS

**A.E. Pismak[a], A.E. Kharitonova[a], E.A. Tsopa[a], S.V. Klimenkov[a]**

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: Evgenij.Tsopa@cs.ifmo.ru

**Article info**
Received 26.01.16, accepted 05.02.16
doi:10.17586/2226-1494-2016-16-2-324-330
Article in Russian

**For citation:** Pismak A.E., Kharitonova A.E., Tsopa E.A., Klimenkov S.V. Evaluation of semantic similarity for sentences in natural language by mathematical statistics methods. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 2, pp. 324–330. doi:10.17586/2226-1494-2016-16-2-324-330

**Abstract**
**Subject of Research.** The paper is focused on Wiktionary articles structural organization in the aspect of its usage as the base for semantic network. Wiktionary community references, article templates and articles markup features are analyzed. The problem of numerical estimation for semantic similarity of structural elements in Wiktionary articles is considered. Analysis of existing software for semantic similarity estimation of such elements is carried out; algorithms of their functioning are studied; their advantages and disadvantages are shown. **Methods.** Mathematical statistics methods were used to analyze Wiktionary articles markup features. The method of semantic similarity computing based on statistics data for compared structural elements was proposed. **Main Results.** We have concluded that there is no possibility for direct use of Wiktionary articles as the source for semantic network. We have proposed to find hidden similarity between article elements, and for that purpose we have developed the algorithm for calculation of confidence coefficients proving that each pair of sentences is semantically near. The research of quantitative and qualitative characteristics for the developed algorithm has shown its major performance advantage over the other existing solutions in the presence of insignificantly higher error rate. **Practical Relevance.** The resulting algorithm may be useful in developing tools for automatic Wiktionary articles parsing. The developed method could be used in computing of semantic similarity for short text fragments in natural language in case of algorithm performance requirements are higher than its accuracy specifications.

**Keywords**
semantic similarity, mathematical statistics, sets, tokens, Wiktionary, semantic analysis, text

## Введение

Одной из проблем в сфере автоматической обработки текстов на естественном языке является отсутствие полноценного словаря, содержащего актуальные, постоянно обновляющиеся данные [1]. Чаще всего такой словарь формируется путем ручного занесения данных [2] и семантических связей [3]. Но этот подход по определению является малоэффективным и весьма затруднительным – создание словаря достаточно большого объема и постоянная актуализация его содержимого потребует огромных временных и финансовых затрат. Таким образом, появляется необходимость в инструментах автоматического формирования и обновления словарей.

В глобальной сети Internet существует множество ресурсов, содержащих данные, на базе которых могут быть построены словари, семантические сети, справочники и тому подобные базы знаний. Одним из источников такого рода данных является открытый словарь Wiktionary [4]. Как и множество других подобных источников, Wiktionary формируется пользователями глобальной сети путем внесения новых данных и редактирования существующих.

Таким образом, Wiktionary является достаточно перспективным источником данных [5] для построения на его основе семантической сети. Однако для этого требуется решить ряд проблем, обусловленных отсутствием в этом словаре жестко заданной структуры статей. Сообщество, которое координирует работу над Wiktionary, выпустило ряд рекомендаций [6], регламентирующих структуру статей и правила их оформления. Тем не менее, эти рекомендации выполняются не всегда, да и сами они содержат ряд недостатков, препятствующих автоматическому извлечению данных из словарных статей [7].

Статья Wiktionary содержит объявление и описание некоторого набора смысловых значений и их свойств [6]. Описание смысловых значений, как правило, задано в виде фразы или предложения на естественном языке. Семантические и грамматические характеристики заданы в виде краткого описания смыслового значения, к которому они относятся, и некоторого набора значений. К характеристикам смыслового значения можно отнести такие данные, как переводы, произношения, синонимы, анаграммы и так далее.

Как уже было упомянуто выше, разметка существенной части статей Wiktionary не соответствует рекомендациям сообщества. Также, вследствие того, что статьи составляются обычными пользователями сети Internet, в большинстве из них присутствуют ошибки; в частности, отсутствуют явные связи между семантико-грамматическими характеристиками и смыслами. Таким образом, для того, чтобы статьи Wiktionary можно было использовать для построения семантической сети, их необходимо привести к единому формату и восстановить отсутствующие связи между смысловыми значениями и их характеристиками (рис. 1). Решение этой задачи сводится к определению номенклатуры типовых шаблонов разметки статей и типовых ошибок, допускаемых пользователями при создании статей. Далее необходимо сформировать связи смысловых значений с их характеристиками. Так как Wiktionary содержит более пяти миллионов смысловых значений, при решении этой задачи необходимо минимизировать затраты вычислительных ресурсов, иначе полный анализ всего словаря займет время, неприемлемое для постоянного поддержания семантической сети в актуальном состоянии.

Задача определения смыслового значения, к которому относится то или иное свойство, сводится к анализу семантической эквивалентности двух заданных предложений или фраз: первое предложение описывает смысловое значение, а второе описывает ту или иную грамматическую или семантическую характеристику, к которому оно относится. Для наглядности рассмотрим фрагмент одной из страниц Wiktionary, в котором представлены четыре синтаксически близких предложения (толкования для глагола «to duck»):

– *To lower the head or body in order to prevent it from being struck by something*;

– *To lower (something) into water*;
– *To go under the surface of water and immediately reappear*;
– *To lower (the head) in order to prevent it from being struck by something*;

и некоторая фраза (ключ соответствия смысловому значению):

– *To lower into the water.*

Необходимо сформировать критерии, определяющие близость заданной фразы к каждому из четырех предложений. Важным условием является обязательное семантическое сходство одного из заданных предложений с фразой.

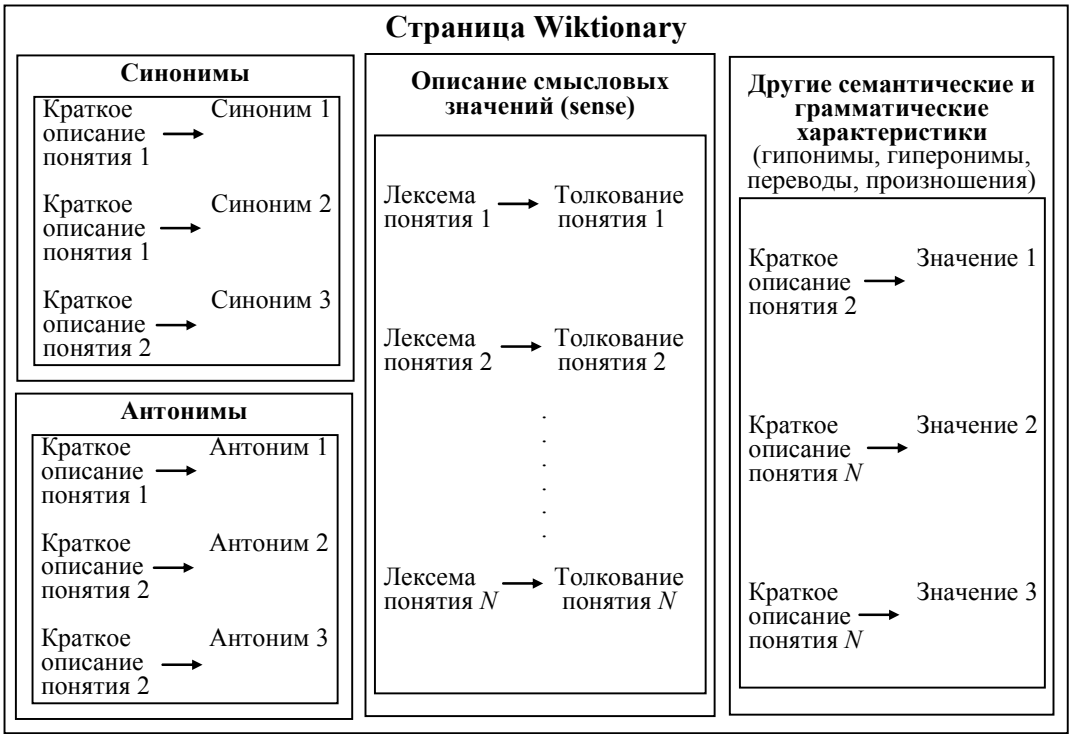


Рис. 1. Структурные элементы статьи Wiktionary

**Обзор существующих решений**

Задача численной оценки семантической близости глубоко исследована, и в настоящий момент существует множество решений, основанных на различных алгоритмах. Такие системы, как Texterra [7], Semanticus [8], S-Space [9], Semantic Vectors [10] и их аналоги используют алгоритм, основанный на семантическом расстоянии [11]. Также существуют системы, предлагающие использовать семантические либо синтаксические анализаторы для построения соответствующих деревьев двух сравниваемых предложений, с последующим анализом и сравнением этих деревьев. Примером такой системы может служить утилита MaltParser [12–14].

Применительно к поставленной задаче использование приведенных выше алгоритмов неприемлемо по затратам на время выполнения. Перечисленные выше существующие продукты решают задачу определения семантической близости за время порядка сотен миллисекунд, а так как Wiktionary содержит более пяти миллионов смысловых значений, то обновление и формирование словаря существующими решениями займет около 25 суток (611 часов).

**Разработка алгоритма оценки семантической близости**

Исходные данные алгоритма представлены в виде ключа и множества $S$, элементы которого описывают смысловые значения (толкования):

$$S\ \{ s_1, s_2, \ldots, s_i \}, \quad (1)$$

где $i$ – это общее количество толкований в статье. Ключ – это предложение или фраза, которая указана для свойства смыслового значения и, как правило, имеет синтаксическую форму, отличную от толкования смыслового значения. Выше уже был представлен пример множества предложений и ключ соответствия этим предложениям. Для каждого элемента из набора $S$ формируется множество токенов $A$, а для ключа – множество токенов $B$:

$$A\ \{ a_1, a_2, \ldots, a_j \}, \quad (2)$$

$$B\ \{ b_1, b_2, \ldots, b_k \}, \quad (3)$$

где $\{ a_1, a_2, \ldots, a_j \}$ – это токены толкования смыслового значения, $\{ b_1, b_2, \ldots, b_k \}$ – токены ключа, $j$ – ко-

личество токенов в предложении $S_i$, $k$ – количество токенов в ключе. Совокупность коэффициентов и параметров исходных множеств может быть охарактеризована множеством коэффициентов $K$:.

$$K\{k_1, k_2, \ldots, k_m\}, \quad (4)$$

где $m$ – это общее количество коэффициентов.

Для множества функций $G$ необходим следующий перечень параметров:

- коэффициент эквивалентности ($E$) – мощность общего подмножества, т.е. количество токенов, которые входят одновременно во множества $A$ и $B$;
- мощность максимального множества ($N_{\max}$) – количество токенов того из двух исходных множеств, где оно максимально;
- мощность минимального множества ($N_{\min}$) – количество токенов того из двух исходных множеств, где оно минимально;
- мощность синтаксического множества ($N_s$) – количество токенов в общем подмножестве, которое находится с учетом последовательностей токенов в исходных множествах.

Некоторый набор численных характеристик не входит в указанный перечень, так как эти характеристики являются зависимыми от исходных параметров. Одна из таких величин – $D$ – отношение мощностей исходных множеств (5):

$$D = \frac{N_{\min}}{N_{\max}}. \quad (5)$$

В первую очередь семантическую близость ключа предложению $S_i$ можно численно охарактеризовать отношением количества токенов в общем синтаксическом подмножестве к количеству токенов в том из исходных множеств, мощность которого максимальна:

$$S_{\max} = \frac{N_s}{N_{\max}}, \quad (6)$$

$$S_{\min} = \frac{N_s}{N_{\min}}. \quad (7)$$

Влияние приведенных выше численных характеристик на коэффициент достоверности было выявлено путем сбора и анализа статистических данных. Примеры собранных статистических данных представлены в табл. 1.

| Ключ | Набор предложений | $S_{\max}$ | $S_{\min}$ | $K_{\text{ord}}$ | $K_{\text{int}}$ | Экспертная оценка |
|---|---|---|---|---|---|---|
| to lower into the water | To lower (something) into water | 0,176 | 0,600 | 0,294 | 4 | есть совпадение |
| | To lower (the head) in order to prevent it from being struck by something | 0,125 | 0,400 | 0,312 | 2 | нет совпадения |
| | The flesh of a duck used as food | 0,111 | 0,200 | 0,555 | 1 | нет совпадения |
| | To lower the head or body in order to prevent it from being struck by something | 0,111 | 0,400 | 0,277 | 3 | нет совпадения |
| a block of solid material | A block of any of various dense materials | 0,333 | 0,600 | 0,555 | 3 | есть совпадение |
| | A small mass of baked dough, especially a thin loaf from unleav ed dough | 0,066 | 0,200 | 0,333 | 2 | нет совпадения |
| | A thin wafer-shaped mass of fried batter | 0,076 | 0,200 | 0,384 | 2 | нет совпадения |
| | A trivially easy task or responsibility | 0,062 | 0,200 | 0,312 | 2 | нет совпадения |

Таблица 1. Пример статистических данных по исходным параметрам

На основе полученного набора данных можно показать, что решающее значение играют параметры $S_{\max}$ и $S_{\min}$. А именно, важно, чтобы какое-либо из представленных значений было максимально. Так как при максимальном $S_{\max}$ параметр $S_{\min}$ может быть минимальным, и наоборот, то за величину $K_{\text{ord}}$,

влияющую на конечный результат примем результат их сложения:

$$K_{\text{ord}} = S_{\max} + S_{\min} \ . \tag{8}$$

Также важной метрикой семантической близости является параметр $K_{\text{int}}$ (9), характеризующий мощность множество, получаемого путем пересечения множеств $A$ и $B$ (т.е. порядок токенов не учитывается).

$$K_{\text{int}} = \left| A \cap B \right| . \tag{9}$$

Коэффициент достоверности – это численная характеристика того факта, что два предложения являются семантически близкими. Для того чтобы снизить влияние погрешностей, периодически возникающих при вычислении $S_{\min}$, $S_{\max}$ и $K_{\text{ord}}$, при расчете коэффициента достоверности необходимо учитывать параметр $N$ (5). В результате проведенных экспериментов было выявлено, что наиболее точное совпадение результатов работы алгоритма с экспертной оценкой получается, если $K_{\text{int}}$ учитывается в формуле расчета коэффициента достоверности (11) следующим образом:

$$K_{\text{corr}} = DK_{\text{int}}^2 \ , \tag{10}$$

где $K_{\text{corr}}$ – значение параметра $K_{\text{int}}$, скорректированное в соответствии с результатами проведенных экспериментов.

Учитывая сказанное выше, формулу для расчета коэффициента достоверности можно записать как

$$R = DK_{\text{int}}^2 (S_{\max} + S_{\min}) = \frac{N_{\min}}{N_{\max}} K_{\text{int}}^2 \left( \frac{N_s}{N_{\max}} + \frac{N_s}{N_{\min}} \right) = K_{\text{int}}^2 \frac{N_s (N_{\max} + N_{\min})}{N_{\max}^2} . \tag{11}$$

Таким образом, общая схема работы алгоритма примет вид, представленный на рис. 2.

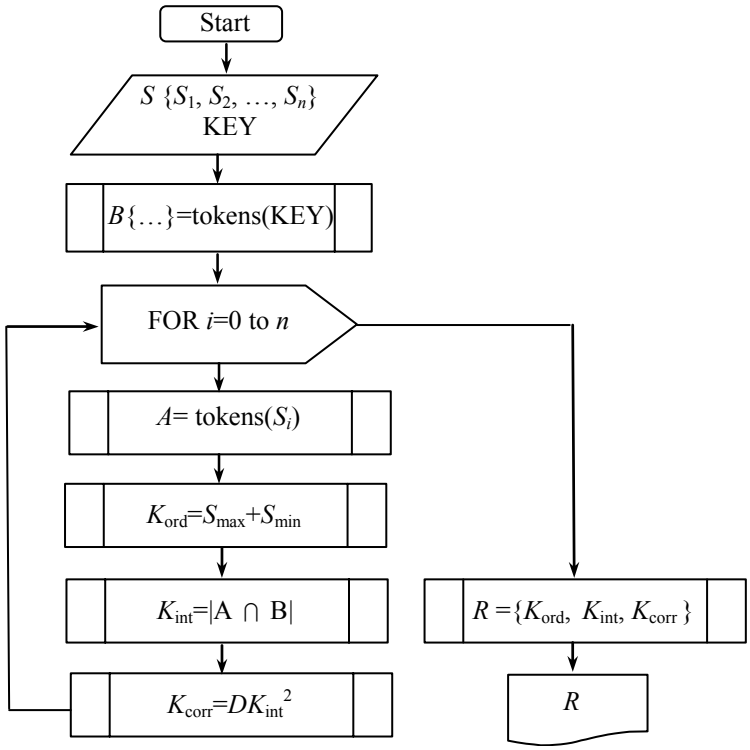


Рис. 2. Блок-схема работы алгоритма

**Результаты экспериментальных исследований**

В табл. 2 представлены результаты работы алгоритма, а именно соответствие значения коэффициента достоверности (9) каждому элементу из набора предложений, сравниваемых по заданному ключу.

Эксперимент был проведен более 9000 раз при разных исходных данных, в результате чего были получены следующие результаты: погрешность алгоритма составляет 9,88%, что вполне приемлемо для решения поставленной задачи. К преимуществам работы алгоритма можно отнести его простоту и быстродействие; время на построение таблицы коэффициентов достоверности для 5 предложений в среднем составляет 1–20 мс на тестовом стенде со следующими характеристиками: CPU Intel Core i7-3612QM 2.10GHz, 8 Gb RAM, OS Linux x64 без виртуализации, с реализацией алгоритма на языке программирования Java и выделенной памяти для виртуальной машины Java 2Gb.

| Ключ | Набор предложений | *R* |
|---|---|---|
| to lower into the water | To lower (something) into water | **3,654** |
| | To lower (the head) in order to prevent it from being struck by something | 0,656 |
| | The flesh of a duck used as food | 0,173 |
| | To lower the head or body in order to prevent it from being struck by something | 1,278 |
| a block of solid material | A block of any of various dense materials | **4,667** |
| | A small mass of baked dough, especially a thin loaf from unleav ed dough | 0,356 |
| | A thin wafer-shaped mass of fried batter | 0,426 |
| | A trivially easy task or responsibility | 0,328 |

Таблица 2. Результаты работы алгоритма

**Заключение**

В результате работы был предложен и реализован алгоритм оценки семантической близости предложений методами математической статистики. Разработанный алгоритм был применен для восстановления связей между смысловыми значениями и их характеристиками в статьях словаря Wiktionary. Экспериментальное исследование показало, что разработанный алгоритм обеспечивает приемлемую погрешность при быстродействии, существенно (более чем на порядок) превышающем скорость существующих решений.

Дальнейшее улучшение алгоритма возможно путем введения дополнительных параметров и корреляционных алгоритмов: расстояние Левенштейна [15], алгоритм нечеткого сравнения токенов, мощности подмножеств для сортированных исходных множеств и так далее. Подобные улучшения могут увеличить корреляцию результатов, но при этом повышая вычислительную сложность алгоритма, что неприемлемо для поставленной задачи.

**Литература**


1. Bessmertny I. Knowledge visualization based on semantic networks // Programming and Computer Software. 2010. V. 6. N 4. P. 197–204. doi: 10.1134/S036176881004002X
2. Nie J.Y., Brisebois M. An inferential approach to information retrieval and its implementation using a manual thesaurus // Artificial Intelligence Review. 1996. V. 10. N 5–6. P. 409–439.
3. Nugumanova A., Bessmertny I. Applying the latent smantic analysis to the issue of automatic extraction of collocations from the domain texts // Communications in Computer and Information Science. 2013. V. 394. P. 92–101. doi: 10.1007/978-3-642-41360-5_8
4. Wiktionary [Электронный ресурс]. Режим доступа: http://wiktionary.org/ свободный. Язык англ. (дата обращения 27.07.2015).
5. Пак А. Парсим русский язык [Электронный ресурс]. Режим доступа: http://habrahabr.ru/post/148124/ свободный. Язык рус. (дата обращения 05.08.2015).
6. Wikipedia: Manual of Style [Электронный ресурс]. Режим доступа: https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style свободный. Язык англ. (дата обращения 12.08.2015).
7. Инновации. Собственные технологии [Электронный ресурс]. Режим доступа: http://www.ispras.ru/technologies/texterra_text_mining_toolkit/ свободный. Язык рус. (дата обращения 01.09.2015).
8. Семантикус Semanticus [Электронный ресурс]. Режим доступа: http://semanticus.ru/ свободный. Язык рус. (дата обращения 08.09.2015).
9. S-Space [Электронный ресурс]. Режим доступа: https://github.com/fozziethebeat/S-Space/wiki/ свободный. Язык англ. (дата обращения 07.09.2015).
10. SemanticVectors [Электронный ресурс]. Режим доступа: https://github.com/semanticvectors/semanticvectors/wiki, свободный. Язык англ. (дата обращения 07.09.2015).
11. Варламов М.И., Коршунов А.В. Расчет семантической близости концептов на основе кратчайших путей в графе ссылок Википедии // Машинное обучение и анализ данных. 2014. Т. 1. № 8. С. 1107–1125.

12. Hall J., Nilsson J., Nivre J. MaltParser [Электронный ресурс]. Режим доступа: http://www.maltparser.org/ свободный. Язык англ. (дата обращения 07.09.2015).
13. Шалиминов И. Методика определения близости на основе синтаксиса [Электронный ресурс]. Режим доступа: https://github.com/ishalyminov/syntactic_classification/wiki, свободный. Язык рус. (дата обращения 03.08.2015).
14. Велихов П.Е. Меры семантической близости статей Википедии и их применение к обработке текстов // Информационные технологии и вычислительные системы. 2009. № 1. С. 23–37.
15. Желудков А.В., Макаров Д.В., Фадеев П.В. Особенности алгоритмов нечеткого поиска // Инженерный вестник. 2014. № 12. С. 501–511.

***Письмак Алексей Евгеньевич*** – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Alexey.Pismak@cs.ifmo.ru

***Харитонова Анастасия Евгеньевна*** – тьютор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Anastassia.Kharitonova@tune-it.ru

***Цопа Евгений Алексеевич*** – ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Evgenij.Tsopa@cs.ifmo.ru

***Клименков Сергей Викторович*** – ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Serge.Klimenkov@cs.ifmo.ru

***Alexey E. Pismak*** – student, ITMO University, Saint Petersburg, 197101, Russian Federation, Alexey.Pismak@cs.ifmo.ru

***Anastassia E. Kharitonova*** – tutor, ITMO University, Saint Petersburg, 197101, Russian Federation, Anastassia.Kharitonova@tune-it.ru

***Evgeniy A. Tsopa*** – assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, Evgenij.Tsopa@cs.ifmo.ru

***Sergey V. Klimenkov*** – assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, Serge.Klimenkov@cs.ifmo.ru