

ИДЕНТИФИКАЦИЯ СМЫСЛОВОЙ БЛИЗОСТИ ФРАГМЕНТОВ ТЕКСТОВ НАУКОМЕТРИЧЕСКИХ БАЗ

Светлана Петрасова, Нина Хайрова, Валерия Киселева

Аннотация: Сложность анализа текстовой информации, содержащейся в наукометрических системах, определяется многозначностью и синонимичностью, которые свойственны языку на всех уровнях его представления, что, прежде всего, влияет на определение смыслового единства языковых единиц. При этом решение задачи усложняется, если речь идет о смысловой близости крупных информационных фрагментов. Поэтому в связи со стремительным ростом объемов информационных ресурсов в наукометрических системах и существующими подходами и методами анализа слабоформализованных данных становятся перспективными задачи обработки текстовой информации на базе интеллектуального анализа. В работе рассматривается информационная технология идентификации смысловой близости фрагментов текстов наукометрических систем. Предложенная технология позволяет определять общие информационные пространства научного взаимодействия авторов за счет идентификации семантически эквивалентных коллокаций в текстах. Технология включает модель формального описания семантико-грамматических характеристик слов атрибутивных, глагольных и субстантивных коллокаций и определение предиката семантической эквивалентности двухсловных коллокаций на основе уравнений алгебры конечных предикатов. Программная имплементация модели представляет собой веб-приложение, определяющее семантически близкие текстовые фрагменты статей, индексируемых в наукометрических базах Google Scholar и Science Direct. В результате определяется эвристическая оценка эффективности разработанной технологии для каждого типа коллокаций.

Ключевые слова: наукометрические базы, смысловая близость фрагментов текстов, информационное пространство, синонимия коллокаций, алгебра конечных предикатов.

ITHEA Keywords: H.3.3 .Information Search and Retrieval, I.2.4. Knowledge Representation Formalisms and Methods

Введение

В последние годы в информационной практике наблюдается возрастающий интерес к наукометрическим исследованиям, обусловленный развитием одноименных баз данных. Современные наукометрические системы формируют статистику, характеризующую динамику показателей востребованности, активности и индексов влияния деятельности ученых. Таким образом, разработка методов структурного анализа современной науки позволяет выявлять исследовательские фронты [King, 2016], ключевые публикации, их авторов, а также мониторить развитие научных направлений и науки в целом.

В настоящее время информационные пространства, представляющие фронты научных исследований, и обозначающие общность, как научных направлений, так и ресурсов, определяются на базе явно выраженных критериев – цитирования, авторства, ключевых слов, а так же критериев коцитирования, проспективных связей и др. Однако в большинстве случаев используемые статистические методы теряют часть знаний об общности информации и являются недостаточными при разработке средств информационного обеспечения библиотек, электронных каталогов, компьютерной библиографии, систем автоматизированного импорта документов и т.п.

Использование технологии идентификации неявно выраженного отношения смысловой близости текстовых фрагментов в работах отдельных авторов позволит выделять единые информационные пространства научных групп в наукометрических системах, что обеспечит релевантный поиск и доступ к научно-исследовательским работам, выполняемым по схожим темам.

Постановка задачи исследования

Целью работы является разработка информационной технологии идентификации смысловой близости текстовых фрагментов в наукометрических системах для определения информационного пространства научного взаимодействия авторов или общих фронтов научных исследований.

Исследованиями научных фронтов или кластеров в науке занимались Ю. Гарфилд [Garfield, 2005], И.В. Маршакова [Маршакова-Шайкевич, 2013], Ю. Грановский [Грановский, 2013] и др.

Среди основных подходов формирования исследовательских фронтов выделяют анализ цитирования документов. Согласно подходу коцитирования [Чайковский та ін, 2013] документы, совместно процитированные в других документах, отражают основные направления современных исследований и создают «ядро» специальности или отрасли науки.

Подобный анализ связей отражен в методе анализа проспективной связи И.В. Маршаковой. В исследовании [Маршакова-Шайкевич, 2013] «близость документов» определяется числом работ, одновременно цитирующих эти документы. Проспективная связь в системе научных публикаций приводит, с одной стороны, к идентификации тематических групп, характеризующих отдельные направления исследуемой области знаний, позволяет проследить их динамику и развитие во времени и объяснить появление новых направлений этой области. С другой стороны, использование этого метода позволяет выявлять научные сообщества (проспективные коллективы). К достоинствам метода проспективной связи относят объективность и точность. К недостаткам метода отнесены трудоемкость его процедур и большое количество операций механического счета [Акоев и др., 2014].

В своих работах [Пенькова и др., 2001], [Евстигнеев, 2004] определяют такие методы статистического анализа документов при выявлении исследовательских фронтов в наукометрических БД как статистический метод, метод подсчета числа публикаций, метод «цитат-индекса». При формировании информационных пространств *статистический метод* использует, кроме показателей количества публикаций, ссылок и ключевых слов, показатели количества ученых, журналов, открытий и др. *Метод подсчета числа публикаций* по научным направлениям дает возможность получить представление об относительном уровне развития отдельных отраслей науки при формировании информационных пространств наукометрических систем. *Метод «цитат-индекса»* базируется на наукометрическими индикаторе – количестве ссылок в научных публикациях.

В работе [Девяткин и др., 2016] для выявления исследовательских фронтов используется гибридная мера близости публикаций. Согласно подходу мера вычисляется на основе трех компонентов: близость по тематическому сходству текстов, при наличии общего цитирования и при наличии общих авторов.

В связи с постоянными изменениями информационного сообщества для адекватного формирования информационных пространств научных сообществ недостаточно использование явно выраженных критериев. Для решения данной задачи необходимо повысить уровень автоматизации обработки естественно-языковой информации, в том числе за счет решения задачи идентификации близких по смыслу фрагментов текстов или словосочетаний.

При определении смысловой близости словосочетаний используются или статистические закономерности, или определяются их синтаксические характеристики. При этом зачастую семантическая информация не учитывается или дополнительно привлекаются тезаурусы.

Среди наиболее разработанных методов определения смысловой близости словосочетаний выделяют:

- метод определения синонимических коллокаций на основе сравнения их переводов. В работе [Hua Wu et al, 2003] кандидаты в синонимические коллокации определяют на основе одноязычного корпуса, а затем, используя их переводы на втором языке, выбирают подходящие пары кандидатов;
- метод выявления перефразирования за счет сходства фрагментов фраз. Метод Pasca и Dienes [Pasca et al, 2005] определяет множество наборов перефразирования из слов и словосочетаний с помощью попарного выравнивания (alignment) небольших фрагментов предложений по множеству предложений веб-документов;
- метод определения сходства контекста на основе анализа параллельных корпусов. Метод Barzilay и McKeown [Barzilay et al, 2001] определяет однословные лексические парафразы, а так же синтаксические парафразы на основе совокупности нескольких английских переводов одного и того же исходного текста.

Все перечисленные подходы работают или на текстах довольно узких предметных областях, или, при статистических подходах, имеют достаточно низкую точность определения эквивалентных словосочетаний. Оба недостатка не позволяют использовать данные подходы при выделении фрагментов информации единых информационных пространств научного взаимодействия авторов в наукометрических системах.

Таким образом, несмотря на достигнутые результаты, на сегодняшний день проблема идентификации смысловой близости фрагментов текста, в частности, в наукометрических системах для определения фронтов научных исследований остается не до конца решенной.

Используемая информационная технология

Разработанная информационная технология [Petrasova et al, 2017] позволяет формировать единые информационные пространства научного взаимодействия авторов (рис. 1).

Технология идентификации смысловой близости текстовых фрагментов включает следующие этапы:

1. Отбор статей наукометрических баз Google Scholar и Science Direct.

На этапах 2, 4 применяется разработанная модель идентификации семантически близких коллокаций [Петрасова и др., 2015] на основе алгебры конечных предикатов.

2. За счет выделения семантико-грамматических характеристик слов коллокаций, рядом стоящих в предложении, идентифицируются субстантивные, атрибутивные и глагольные коллокации.

3. Использование WordNet¹ для определения синонимичных слов коллокаций.

¹ <https://wordnet.princeton.edu/>

WordNet позволяет выделять классы эквивалентности (синсеты), то есть классы синонимических в каком из своих смыслов терминов, содержащиеся в текстах статей.

4. Далее определяется семантическая близость выделенных фрагментов.

Синонимичные слова могут образовывать близкие по смыслу словосочетания, например, «хранить данные» и «содержать сведения», в то же время, могут формировать близкие по смыслу словосочетания, «хранение данных» \neq «информация репозитария».

Итак, коллокации могут считаться близкими по смыслу, если семантико-грамматические характеристики коллокатов словосочетаний удовлетворяют предикату семантической эквивалентности [Petrasova et al, 2015].

5. На последнем этапе разработанной технологии, используя найденные подобные эквивалентные фрагменты информации, выделяем единое информационное пространство научного взаимодействия авторов.

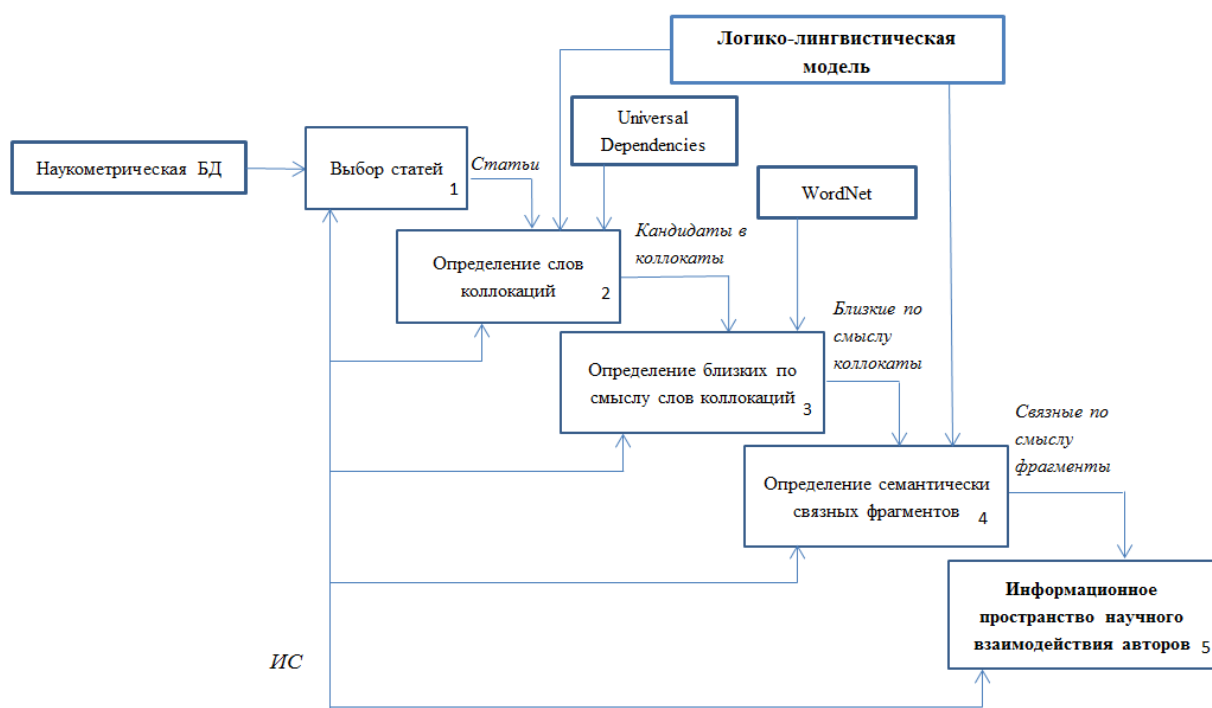


Рис. 1. Основные этапы разработанной информационной технологии

Описание модели

Идентификация информационно-лингвистических сущностей, в частности коллокации и отношений, с характерными для естественного языка гибкостью и многозначностью требует интеллектуальных средств обработки естественно-языковых текстов.

В качестве формального аппарата для построения модели идентификации дискретного, конечного набора смысловых сущностей и отношений в текстовой информации наукометрических систем использовался аппарат алгебры конечных предикатов.

Для описания семантических и грамматических характеристик фрагментов информации, а именно слов коллокаций, были введены предметные переменные, определяющие:

- часть речи: a^N (noun), a^A (adjective), a^V (verb);
- синтаксическую роль существительного: a^{NSub} (subject), a^{NObj} (object) и прилагательного: a^{AAtt} (attribute), a^{APr} (predicative);
- транзитивность глагола: a^{VTr} (transitive), a^{VIntr} (intransitive);
- возможные семантические роли существительных: c^{Ag} – агент (активный участник ситуации), c^{Att} – атрибут (связь предмета и признака), c^{Pac} – пациент (пассивный участник ситуации или объект действия), c^{Adr} – адресат (получатель сообщения), c^{Ins} – инструмент (участник, с помощью которого осуществляется действие, или инструмент осуществления действия), c^M – место (местонахождение одного из участников ситуации).

Введенный на множестве словоформ предикат $P(x)$ превращается в 1, если главная словоформа словосочетаний x имеет определенную семантико-грамматическую информацию. Множество допустимых семантико-грамматических характеристик зависимого слова словосочетания y описывается предикатом $P(y)$.

Предикат идентификации рядом стоящих сущностей, образующих коллокации английского языка:

$$P(x, y) = (x^{NSubAg} \vee x^{NSubOfAg} \vee x^{VTr})(y^{NObjAtt} \vee y^{NObjPac} \vee y^{AAtt} \vee y^{APr})$$

Введенный предикат семантической эквивалентности между коллокациями определяет семантико-грамматические характеристики коллокатов близких по смыслу словосочетаний. Отношение семантической эквивалентности двух двухсловных коллокаций может быть определено как:

$$P(x_1, y_1) \times P(x_2, y_2) = \gamma_i(x_1, y_1, x_2, y_2) \wedge P(x_1, y_1) \wedge P(x_2, y_2)$$

где знак \times обозначает операцию определения смысловой близости, знак \wedge определяет конъюнкцию, предикат $\gamma_i(x_1, y_1, x_2, y_2)$ исключает коллокации, между которыми не может быть установлена смысловая эквивалентность.

Предикат $\gamma_i(x_1, y_1, x_2, y_2) = x_1^{VTr} y_1^{NObjPac} x_2^{VTr} y_2^{NObjPac}$ показывает семантическую близость глагольных коллокаций ($V_x N_y$), например, identify information \approx extract data.

Предикат $\gamma_2(x_1, y_1, x_2, y_2) = x_1^{NSubOfAg} y_1^{NObjAtt} y_2^{NObjAtt} x_2^{NSubAg} \vee x_1^{NSubOfAg} y_1^{NObjAtt} x_2^{NSubOfAg} y_2^{NObjAtt} \vee y_1^{NObjAtt} x_1^{NSubAg} y_2^{NObjAtt} x_2^{NSubAg}$ показывает семантическую близость субстантивных коллокаций ($N_x N_y$), например, figure means \approx pattern technique.

Предикат $\gamma_3(x_1, y_1, x_2, y_2) = y_1^{AAtt} x_1^{NSubAg} x_2^{NSubAg} y_2^{APr} \vee y_1^{AAtt} x_1^{NSubAg} y_2^{AAtt} x_2^{NSubAg} \vee x_1^{NSubAg} y_1^{APr} x_2^{NSubAg} y_2^{APr}$ показывает семантическую близость между атрибутивными коллокациями ($A_y N_x$), например, important topics \approx essential issues, key field \approx central area.

Программная реализация

Программная имплементация модели представляет собой веб-приложение, анализирующее текстовую информацию для идентификации семантически близких текстовых фрагментов, а именно коллокаций (рис. 2).

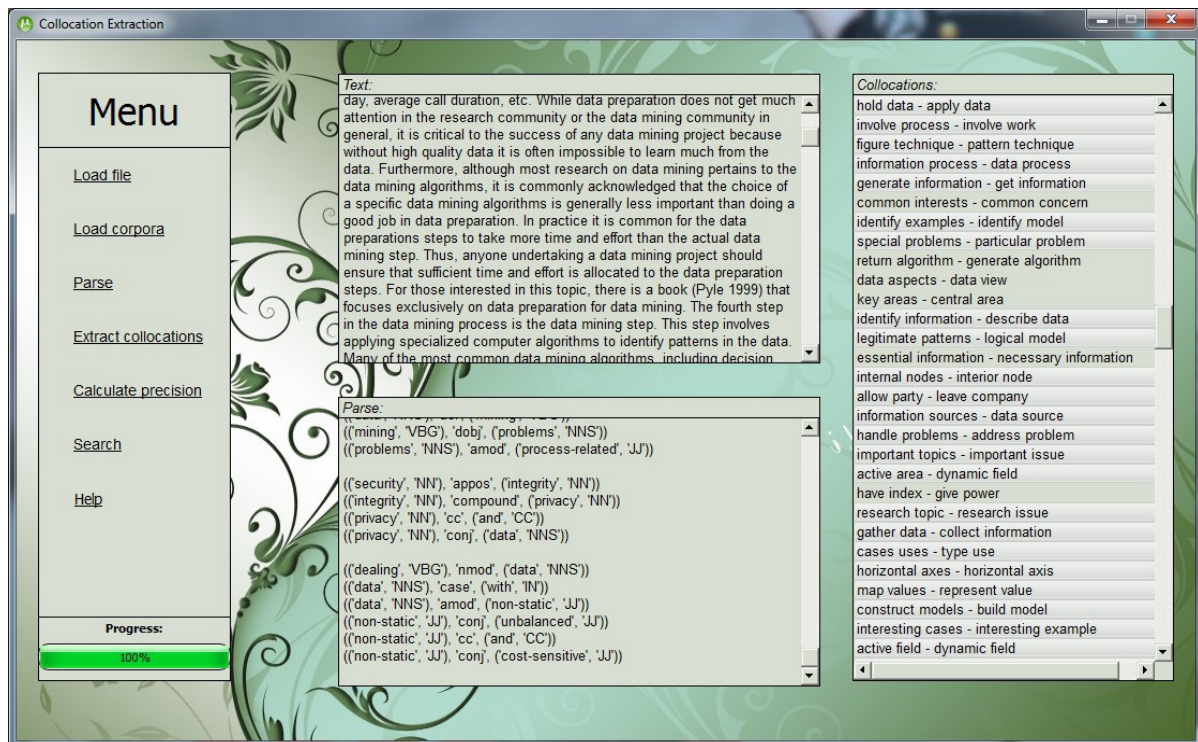


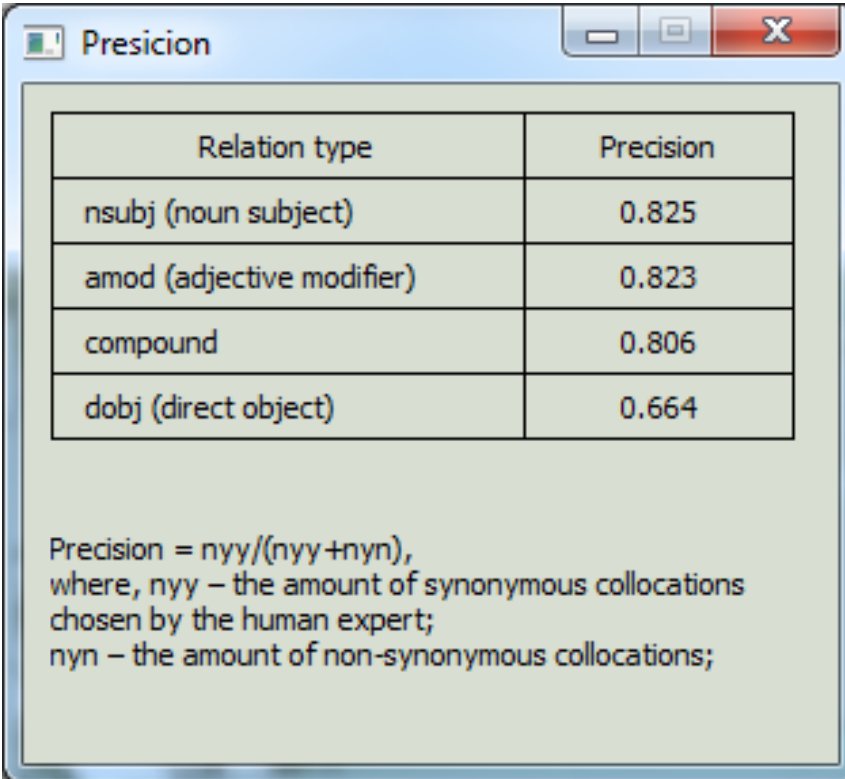
Рис. 2. Программная имплементация модели идентификации близких по смыслу коллокаций

Программа отображает извлеченную информацию в виде:

- перечня синонимичных коллокаций. Для извлечения кандидатов в коллокаты используются синсеты WordNet, содержащие синонимы слов коллокаций, которые отсортированы по частям речи и значениям;

- грамматической характеристики слов коллокаций, полученной с помощью модуля Universal Dependencies². В атрибутивных, глагольных и субстантивных коллокациях Universal Dependencies определяет такие типы связей, как *amod: adjectival modifier*, *dobj: direct object*, *compound: compound*, *nsubj: nominal subject*;
- текстовой информации или ее источников, т.е. статей, индексируемых в Google Scholar и ScienceDirect, из которых данные коллокации были извлечены.

Для оценки эффективности работы технологии в меню программы было создано пункт Calculate Precision (рис. 3). Метрика Precision рассчитывается по формуле $Precision = n_{yy} / (n_{yy} + n_{yn})$.



Relation type	Precision
nsubj (noun subject)	0.825
amod (adjective modifier)	0.823
compound	0.806
dobj (direct object)	0.664

Precision = $n_{yy} / (n_{yy} + n_{yn})$,
 where, n_{yy} – the amount of synonymous collocations
 chosen by the human expert;
 n_{yn} – the amount of non-synonymous collocations;

Рис. 3. Результаты Calculate Precision

Выводы

Результатом данного исследования является разработка технологии идентификации близких по смыслу фрагментов текстовой информации в наукометрических системах. Программная реализация разработанной модели идентификации близких по смыслу коллокаций, основывающейся на использовании алгебры конечных предикатов, позволяет определить информационные пространства научного взаимодействия авторов статей.

² <http://universaldependencies.org/u/dep/index.html>

Литература

- [Barzilay et al, 2001] Barzilay R., McKeown K.R. Extracting Paraphrases from a Parallel Corpus. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL '01). Stroudsburg, PA, USA, 2001. P. 50–57.
- [Garfield, 2005] Garfield E. The Agony and the Ecstasy – The History and the Meaning of the Journal Impact Factor. In Proc. of Inter. Congress on Peer Review and Biomedical Publication. Chicago, 2005. Access mode: <http://garfield.library.upenn.edu/papers/jifchicago2005.pdf>.
- [Hua Wu et al, 2003] Hua Wu, Ming Zhou Synonymous Collocation Extraction Using Translation Information. In Proceedings of the 41th Annual Meeting on Association for Computational Linguistics (ACL '03). Stroudsburg, PA, USA, 2003. Vol. 1. P. 120–127.
- [King, 2016] King C. Research Fronts 2016: The Hottest Areas in Science. Access mode: <http://stateofinnovation.com/research-fronts-2016-the-hottest-areas-in-science>
- [Pasca et al, 2005] Pasca M., Dienes P. Aligning Needles in a Haystack: Paraphrase Acquisition across the Web. In Proceedings of the Second International Joint Conference: Natural Language Processing (IJCNLP 2005). Korea, 2005. P. 119–130.
- [Petrasova et al, 2015] Petrasova S. and Khairova N. Automatic Identification of Collocation Similarity. In Proceedings of 10th International Scientific and Technical Conference: Computer Science & Information Technologies (CSIT'2015), Lviv, 2015. P. 136–138.
- [Petrasova et al, 2017] Petrasova S. and Khairova N. Using a Technology for Identification of Semantically Connected Text Elements to Determine a Common Information Space. Cybernetics and Systems Analysis, Springer. Vol. 53 (1). 2017. P. 115–124.
- [Акоев и др., 2014] Акоев М.А., Маркусова В.А. и др. Руководство по наукометрии: индикаторы развития науки и технологии : монография. – Екатеринбург: Изд-во Урал. ун-та, 2014. – 250 с.
- [Грановский, 2013] Грановский Ю.В. Наукометрия и управление научными коллективами // Наукоевческие исследования. – 2013. – С. 127–150.
- [Девяткин и др., 2016] Девяткин Д.А., Швец. А.В., Тихомиров И.А. Выявление направлений исследований и научных коллективов на основе анализа полнотекстовых коллекций научных публикаций. – Режим доступа : <http://www.gpntb.ru/win/inter-events/crimea2016/disk/2046.pdf>
- [Евстигнеев, 2004] Евстигнеев В.А. Наукометрические исследования в информатике // Новосибирская школа программирования. Переключка времен : сб. тр. – Новосибирск: Ин-т систем информатики им. А.П. Ершова СО РАН, 2004. – С. 203–217.
- [Петрасова и др., 2015] Петрасова С.В., Хайрова Н.Ф. Логико-лингвистическая модель идентификации семантически эквивалентных коллокаций // Вестник Национального

технического университета «Харьковский политехнический институт». – Харьков : НТУ «ХПИ», 2015. – № 58 (1167). – С. 14–17.

[Чайковський та ін., 2013] Чайковський Ю.Б., Сіліна Ю.В., Потоцька О.Ю. Наукометричні бази та їх кількісні показники (Частина I. Порівняльна характеристика наукометричних баз) // Вісник Національної академії наук України. – 2013. – № 8. – С. 89–98.

[Маршакова-Шайкевич, 2013] Маршакова-Шайкевич И.В. Роль библиометрии в оценке исследовательской активности науки // УБС. – 2013. – № 44. – С. 210–247.

[Пенькова и др., 2001] Пенькова О.В., Тютюнник В.М. Информетрия, наукометрия и библиометрия: наукометрический анализ современного состояния // Вестник ТГУ. – 2001. – Т. 6. – № 1. – С. 86–88.

Authors' Information



Светлана Петрасова – к.т.н., старший преподаватель кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Кирпичева, 2, Харьков, 61002, Украина; e-mail: svetapetrasova@gmail.com

Научные интересы: искусственный интеллект, компьютерная лингвистика, Information Extraction, Natural Language Processing



Нина Хайрова – профессор, д.т.н., профессор кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Кирпичева, 2, Харьков, 61002, Украина; e-mail: nina_khajrova@yahoo.com

Научные интересы: искусственный интеллект, компьютерная лингвистика, экстракция знаний из текстов, Text Mining, Opinion Mining, Web Mining, Natural Language Processing



Валерия Киселева – магистр Национального технического университета «Харьковский политехнический институт», ул. Кирпичева, 2, Харьков, 61002, Украина; e-mail: lerakiseleva@yahoo.com

Научные интересы: компьютерная лингвистика, идентификация семантической синонимии коротких фрагментов текстов.

Identification of Semantic Similarity of Text Fragments in Scientometric Bases

Svitlana Petrasova, Nina Khairova, Valeriia Kysilova

Abstract: *This paper considers the information technology for identification of semantic similarity of text fragments in scientometric systems. The proposed technology allows determining common information spaces of authors' scientific interaction due to identification of semantic equivalence of collocations in texts. The technology includes a model for a formal description of the semantic and grammatical characteristics of words in attributive, verbal and substantive collocations and identification of the semantic equivalence predicate for two-word collocations based on the algebra of finite predicates. The developed software implementation is a web application that defines semantically connected text fragments of articles indexed in Google Scholar and Science Direct. As a result, an effectiveness estimate of the developed technology for each type of collocations is determined.*

Keywords: *scientometric bases, semantic similarity of text fragments, information space, collocation synonymy, algebra of finite predicates.*