Вычисление меры семантической связности на основе вики проектов

**Человек:** Объектом исследования данной работы является мера семантической связности текстов, предметом исследования - алгоритм вычисления меры семантической связности текстов. В статье особое внимание уделяется методу определения гибридной меры семантической связности двух понятий. Этот метод лежит в основе алгоритма вычисления сходства двух текстов. В качестве источников знаний используются Вики-проекты: Википедия и Викисловарь. Их совместное использование позволяет охватить значительно большее количество слов, по сравнению с использованием одного из Вики-проектов. В методе используется известная мера Wikisim. Эта мера простая, но имеет хорошую производительность. В классическом методе Wikisim используется только Википедия, поэтому он адаптирован и для Викисловаря. Методология проведения работы основана на моделировании процесса извлечения высококачественных знаний из Вики-проектов – индивидуальной работы независимых добровольцев. Новизна исследований заключается в комбинировании двух источников знаний Википедии и Викисловаря и создании на их основе новой гибридной меры семантической связанности понятий. Основной вывод работы состоит в том, что совмещение формальных (Викисловарь) и неформальных (Википедия) источников знаний может привести к более качественной оценке семантической связности между текстовыми единицами. Описанный метод может применяться в экономики, социологии и политике для выяснения мнения людей по интересующим вопросам.

**Key words:** слово, текст, Викисловарь, Википедия, вики проект, источник знаний, гибридная мера, семантическая связность, понятие, частотная характеристика

=================================

**FastText\_KMeans\_Clean:** При изучении работ по данной теме, было выяснено, что методы, основанные использовании Википедии, хорошо вычисляют меры семантической связности в предметных областях и имен собственных. Всего слов в тексте. Слова проранжированы по частоте встречаемости и первые пять понятий с наибольшими значениями показаны в табл.2. Таблица 2 – Частотные характеристики. WikiSim on Wikipedia. N пары концептов, l. Ядром алгоритма вычисления меры семантической связности двух текстов является гибридная мера семантической связности двух понятий, основанная на применении метода WikiSim.

**Key words part:** 0.7916666666666666

=================================

**FastText\_KMeans\_Raw/:** Но было также установлено, что Википедию невозможно применить для анализа связности глаголов, прилагательных и остальных частей речи, кроме существительных, в связи с тем, что Википедия делает акцент на энциклопедические термины [4]. Слова проранжированы по частоте встречаемости и первые пять понятий с наибольшими значениями показаны в табл.2. Таблица 2 – Частотные характеристики. WikiSim on Wikipedia. N пары концептов, l. Ядром алгоритма вычисления меры семантической связности двух текстов является гибридная мера семантической связности двух понятий, основанная на применении метода WikiSim.

**Key words part:** 0.7916666666666666

=================================

**FastText\_PageRank\_Clean/:** Сталлоне открыл памятную доску у монумента Рокки (https://ria.ru/culture/20180407/1518120151.html). Количество (частота встречаемости), f. Результаты показаны в табл. WikiSim on Wikipedia. WikiSim on Wiktionary. l – порядковый номер пары концептов ;. N пары концептов, l. N пары концептов, l.

**Key words part:** 0.375

=================================

**FastText\_PageRank\_Raw/:** Сталлоне открыл памятную доску у монумента Рокки (https://ria.ru/culture/20180407/1518120151.html). Количество (частота встречаемости), f. Для этого выполняется последовательный анализ пар статей: (, ), (, ), (, ). WikiSim on Wikipedia. WikiSim on Wiktionary. l – порядковый номер пары концептов ;. N пары концептов, l. N пары концептов, l.

**Key words part:** 0.375

=================================

**Mixed\_ML\_TR/:** В противоположность этому, охват WordNet ограничен и почти не покрывает конкретную предметную область и обладает скудным охватом имен собственных [10]. % в тексте (относительная частота), β. Для этого выполняется последовательный анализ пар статей: (, ), (, ), (, ). Выполнив анализ множества пар понятий по методу WikiSim on Wikipedia переходим к рассмотрению этого множества по методу WikiSim on Wiktionary . 3. Таблица 3 – Меры семантической связности концептов первых двух текстов. 4. Таблица 4 – Гибридная мера семантической связности концептов первого и второго текстов. Аналогичным образом находятся меры семантической связности концептов для пар текстов (t 1, t 3), (t 2, t 3). Ядром алгоритма вычисления меры семантической связности двух текстов является гибридная мера семантической связности двух понятий, основанная на применении метода WikiSim. Благодаря быстрому и общедоступному развитию вики проектов, созданию алгоритма вычисления меры семантической связности текстов мы можем получать данные, соответствующие современным и актуальным на сегодняшний день реалиям.

**Key words part:** 0.7083333333333334

=================================

**MultiLingual\_KMeans/:** В противоположность этому, охват WordNet ограничен и почти не покрывает конкретную предметную область и обладает скудным охватом имен собственных [10]. % в тексте (относительная частота), β. Для этого выполняется последовательный анализ пар статей: (, ), (, ), (, ). Выполнив анализ множества пар понятий по методу WikiSim on Wikipedia переходим к рассмотрению этого множества по методу WikiSim on Wiktionary . Аналогичным образом находятся меры семантической связности концептов для пар текстов (t 1, t 3), (t 2, t 3). Благодаря быстрому и общедоступному развитию вики проектов, созданию алгоритма вычисления меры семантической связности текстов мы можем получать данные, соответствующие современным и актуальным на сегодняшний день реалиям.

**Key words part:** 0.6666666666666666

=================================

**Multilingual\_PageRank/:** Стремительный рост социальных медиа вызывают большой интерес к приложениям, выполняющим анализ блогосферы. Блог-постам характерны тексты небольшого объема. Это многоязычный, и состоящий примерно из 3,5 млн записей. Но в нем отсутствуют термины предметных областей и имен собственных. Комиссия США по торговле одобрила введение пошлин на поставки алюминиевой фольги из КНР (https://www.kommersant.ru/doc/3598977). США сообщили о возможном введении пошлин на поставки фольги из Китая (https://www.rbc.ru/rbcfreenews/59f3ee139a7947a5db3a3cd4). В настоящей работе мы ограничились мощностью множества концептов текста равной пяти. Так как все тексты небольшого объема, то остальные концепты встречались не более одного раза.

**Key words part:** 0.4166666666666667

=================================

**RuBERT\_KMeans\_Without\_ST/:** По сравнению со стандартным словарем, таким как Oxford English Dictionary, Wiktionary предлагает широкий спектр семантических и лексических отношений, следовательно, его можно назвать тезаурусом [11]. 50. . На первом этапе находятся частотные характеристики: абсолютная и относительные частоты встречаемости слов в тексте fij (i – индекс слова, j – индекс текста). N пары концептов, l. 5. Таблица 5 – Результаты расчета меры семантической связности двух текстов ,.

**Key words part:** 0.6666666666666666

=================================

**RuBERT\_KMeans\_With\_ST/:** По сравнению со стандартным словарем, таким как Oxford English Dictionary, Wiktionary предлагает широкий спектр семантических и лексических отношений, следовательно, его можно назвать тезаурусом [11]. - вычисление меры семантической связности двух концептов с использованием метода WikiSim и источника знаний Wikipedia, условно обозначим его как "WikiSim on Wikipedia";. 3) вычисление меры семантической связности двух текстов. 3. Таблица 3 – Меры семантической связности концептов первых двух текстов. Результаты вычислений показаны в табл. N пары концептов, l. где и относительные частоты встречаемости понятий первого и второго текста соответственно (см. табл.

**Key words part:** 0.6666666666666666

=================================

**RUBERT\_page\_rank\_Without\_ST/:** Блог-постам характерны тексты небольшого объема. Но в нем отсутствуют термины предметных областей и имен собственных. % в ядре (относительная частота), α. % в тексте (относительная частота), β. и – весовые коэффициенты, позволяющие учесть частоту встречаемости понятий первого и второго текстов соответственно.

**Key words part:** 0.4583333333333333

=================================

**RUBERT\_page\_rank\_With\_ST/:** Стремительный рост социальных медиа вызывают большой интерес к приложениям, выполняющим анализ блогосферы. США сообщили о возможном введении пошлин на поставки фольги из Китая (https://www.rbc.ru/rbcfreenews/59f3ee139a7947a5db3a3cd4). Сталлоне открыл памятную доску у монумента Рокки (https://ria.ru/culture/20180407/1518120151.html). Для этого выполняется последовательный анализ пар статей: (, ), (, ), (, ). весовые коэффициенты и вычисляются по формулам.

**Key words part:** 0.375

=================================

**RUSBERT\_KMeans\_Without\_ST/:** Семантическая близость состоит из семантических связей между двумя терминами, которые имеют аналогичную природу, состав и атрибуты. 5. Таблица 5 – Результаты расчета меры семантической связности двух текстов ,. В статье не разделялись вики проекты на англоязычные и русскоязычные так как, если слова нет в русскоязычном Вики-проекте можно использовать его англоязычный аналог. Совмещение формальных (структурированных – Викисловарь) и неформальных (неструктурированных - Википедия) источников знаний может привести к качественной оценке семантической связности между текстовыми единицами.

**Key words part:** 0.8333333333333334

=================================

**RUSBERT\_KMeans\_With\_ST/:** 50. . На первом этапе находятся частотные характеристики: абсолютная и относительные частоты встречаемости слов в тексте fij (i – индекс слова, j – индекс текста). В данной статье не описывается процесс построения пространственного графа, т.к. принцип его описание построения примерно аналогичен описанию, приведенному в работе [2]. По формуле (1) вычисляются меры семантической связности концептов . Выполнив анализ множества пар понятий по методу WikiSim on Wikipedia переходим к рассмотрению этого множества по методу WikiSim on Wiktionary . где и относительные частоты встречаемости понятий первого и второго текста соответственно (см. табл.

**Key words part:** 0.7083333333333334

=================================

**RUSBERT\_page\_rank\_Without\_ST/:** Блог-постам характерны тексты небольшого объема. Но в нем отсутствуют термины предметных областей и имен собственных. 1) нахождение частотных характеристик;. N пары концептов, l. N пары концептов, l.

**Key words part:** 0.5

=================================

**RUSBERT\_page\_rank\_With\_ST/:** 1) нахождение частотных характеристик;. Всего слов в тексте. % в ядре (относительная частота), α. N пары концептов, l. N пары концептов, l.

**Key words part:** 0.5416666666666666

=================================

**Simple\_PageRank/:** Она охватывает не только семантическую близость, но также концепты, не имеющие явную схожую природу, состав и атрибуты, но тесно связанные. При изучении работ по данной теме, было выяснено, что методы, основанные использовании Википедии, хорошо вычисляют меры семантической связности в предметных областях и имен собственных. Но было также установлено, что Википедию невозможно применить для анализа связности глаголов, прилагательных и остальных частей речи, кроме существительных, в связи с тем, что Википедия делает акцент на энциклопедические термины [4]. 50. . На первом этапе находятся частотные характеристики: абсолютная и относительные частоты встречаемости слов в тексте fij (i – индекс слова, j – индекс текста). 3. Таблица 3 – Меры семантической связности концептов первых двух текстов. Благодаря быстрому и общедоступному развитию вики проектов, созданию алгоритма вычисления меры семантической связности текстов мы можем получать данные, соответствующие современным и актуальным на сегодняшний день реалиям.

**Key words part:** 0.7916666666666666

=================================

**TextRank/:** - вычисление меры семантической связности двух концептов с использованием метода WikiSim и источника знаний Wikipedia, условно обозначим его как "WikiSim on Wikipedia";. - вычисление меры семантической связности двух концептов с использованием метода WikiSim и источника знаний Wiktionary, условно обозначим его как "WikiSim on Wiktionary";. 3. Таблица 3 – Меры семантической связности концептов первых двух текстов. 4. Таблица 4 – Гибридная мера семантической связности концептов первого и второго текстов. Результаты вычисления меры семантической связности двух текстов показаны в табл. Ядром алгоритма вычисления меры семантической связности двух текстов является гибридная мера семантической связности двух понятий, основанная на применении метода WikiSim.

**Key words part:** 0.7083333333333334

=================================

**TF-IDF\_KMeans/:** Семантические меры используются в качестве основных компонентов в большом количестве приложений, которые очень сильно зависят от оценок семантических ассоциаций. Википедия имеет преимущества по сравнению с другими источниками знаний, такими как WordNet и Wiktionary. Для каждого слова он имеет страницу статьи, в которой перечисляются различные классы слов. Мера семантической связности двух текстов. В результате сравнения определяется множество совпавших ссылок. Результаты вычислений показаны в табл. N пары концептов, l. где и относительные частоты встречаемости понятий первого и второго текста соответственно (см. табл.

**Key words part:** 0.75

=================================

**Текст:** Вычисление меры семантической связности на основе вики проектов. Введение. Семантические меры используются в качестве основных компонентов в большом количестве приложений, которые очень сильно зависят от оценок семантических ассоциаций. Сфера применения семантических мер многопрофильна, начиная от компьютерной лингвистики до искусственного интеллекта и от когнитивной психологии до восстановления информации.. В области обработки естественно-языковой информации более широко распространен термин «семантическая близость», чем «семантическая связность». Семантическая близость состоит из семантических связей между двумя терминами, которые имеют аналогичную природу, состав и атрибуты. Примерами семантической близости являются отношения синонимии, гиперонимии и гипонимии. В работе [6] семантическая близость определена как таксономическая близость двух слов. Семантическая связность тесно связана с семантической близостью, но является более общим термином, включая множество классических и неклассических отношений [5]. Она охватывает не только семантическую близость, но также концепты, не имеющие явную схожую природу, состав и атрибуты, но тесно связанные.. Множество приложений искусственного интеллекта, по сути, основаны на мерах семантических ассоциаций. Добыча мнения, также известна как анализ настроений, является задачей автоматического определения отношения (мнения, оценки, эмоций) людей в отношении объектов и их атрибутов [7]. Стремительный рост социальных медиа вызывают большой интерес к приложениям, выполняющим анализ блогосферы. Блог-постам характерны тексты небольшого объема. Поэтому стандартные методы определения семантической ассоциации текстов блог-постов не совсем подходят, т.к. таким текстам свойственны нестрогие грамматические структуры.. Данная работа посвящена методу определения меры семантической связности текстов на основе вики проектов, позволяющему эффективно определять меру сходства двух текстов небольшого объема. Предлагаемый метод основан на методе Wikisim [3]. Этот метод использует в качестве источника знаний "Википедию". В работе предлагается добавить новый источник знаний "Викисловарь".. . Обоснование выбора источников знаний. Википедия является совместно построенной, многоязычной и свободно доступной онлайн-энциклопедией [1, 2]. Википедия имеет преимущества по сравнению с другими источниками знаний, такими как WordNet и Wiktionary. Наиболее важным из всех является ее превосходное освещение концептов, особенно имен собственных. В ней описано огромное количество знаний, связанных с конкретной предметной областью, что делает её привлекательным ресурсом. В работе [9] было проведено исследование с целью изучения охвата Википедии в области продовольствия и сельского хозяйства. Они показали, что Википедия обеспечивает хороший охват сельскохозяйственных тем, приближающийся к охвату профессионального тезауруса.. В противоположность этому, охват WordNet ограничен и почти не покрывает конкретную предметную область и обладает скудным охватом имен собственных [10]. Викислова́рь (англ. Wiktionary) – свободно пополняемый многофункциональный многоязычный словарь и тезаурус. Словарь Wiktionary является совместно построенным и доступным онлайн словарем. Это многоязычный, и состоящий примерно из 3,5 млн записей. По сравнению со стандартным словарем, таким как Oxford English Dictionary, Wiktionary предлагает широкий спектр семантических и лексических отношений, следовательно, его можно назвать тезаурусом [11].. Wiktionary имеет много общих черт с WordNet. Для каждого слова он имеет страницу статьи, в которой перечисляются различные классы слов. Каждый класс слова соответствует концепту. Следуя WordNet, Wiktionary также определяет лексические, семантические отношения, такие как части речи, произношение, синонимы, гиперонимы, гипонимы, перевод на другие языки.. При изучении работ по данной теме, было выяснено, что методы, основанные использовании Википедии, хорошо вычисляют меры семантической связности в предметных областях и имен собственных. Но было также установлено, что Википедию невозможно применить для анализа связности глаголов, прилагательных и остальных частей речи, кроме существительных, в связи с тем, что Википедия делает акцент на энциклопедические термины [4].. В этих случаях удобно применять Викисловарь, в котором мы можем установить явные морфологические, синтаксические и семантические свойства. Но в нем отсутствуют термины предметных областей и имен собственных.. Плюсы одного вики проекта взаимно покрывают минусы другого, что говорит об эффективности меры, базирующейся на двух источников знаний. Именно этот факт привел к решению создания гибридной меры семантической связности концептов.. . Мера семантической связности двух текстов. Мера основана на комбинировании аспектов проектов Википедии и Викисловаря, как источников знаний. Существующие меры семантической связности используют основную структуру Википедии как неформального источника знаний для вычисления семантических ассоциаций, а именно сеть гиперссылок. Гиперссылки Википедии являются связью между двумя статьями, разделяющими некоторый контекст. Статьи, имеющие ссылки на конкретную статью называются статьями in-links. Аналогичным образом, статьи, которые упоминаются в анализируемой статье называются статьями out-links.. Описываемая мера основана на методе WikiSim, который показывает лучшие результаты по сравнению с другими известными методами [4]. WikiSim учитывает долю общих ссылок двух статей и , двух анализируемых концептов и .. Подход начинается с согласования входных понятий и к их соответствующим статьям Википедии и соответственно.. Установленное для , множество ссылок (), состоящее из ее отдельных статей in-link и out-link, сравнивается с множеством ссылок () статьи . В результате сравнения определяется множество совпавших ссылок. Затем множество совпавших ссылок используется для вычисления меры связности.. (1). В работе предложен следующий алгоритм вычисления меры семантической связности двух текстов:. 1) нахождение частотных характеристик;. 2) нахождение меры семантической связности двух концептов:. - вычисление меры семантической связности двух концептов с использованием метода WikiSim и источника знаний Wikipedia, условно обозначим его как «WikiSim on Wikipedia»;. - вычисление меры семантической связности двух концептов с использованием метода WikiSim и источника знаний Wiktionary, условно обозначим его как «WikiSim on Wiktionary»;. - вычисление гибридной меры семантической связности двух концептов (hybrid measure).. 3) вычисление меры семантической связности двух текстов.. Рассмотрим описание алгоритма на конкретном примере.. Для описания работы алгоритма были выбраны три статьи, характеристики текстов статей, полученные после предварительной обработки текстов, представлены в табл.1. Таблица 1 – Исходные данные. №. Заголовок статьи. Всего слов в тексте. Слов в тексте после удаления стоп-слов. Слов в ядре текста. 1. 2. 3. 4. 5. 1. Комиссия США по торговле одобрила введение пошлин на поставки алюминиевой фольги из КНР (https://www.kommersant.ru/doc/3598977). 107. 73. 57. 2. США сообщили о возможном введении пошлин на поставки фольги из Китая (https://www.rbc.ru/rbcfreenews/59f3ee139a7947a5db3a3cd4). 158. 99. 67. 3. Сталлоне открыл памятную доску у монумента Рокки (https://ria.ru/culture/20180407/1518120151.html). 103. 75. 50. . На первом этапе находятся частотные характеристики: абсолютная и относительные частоты встречаемости слов в тексте fij (i – индекс слова, j – индекс текста). Слова проранжированы по частоте встречаемости и первые пять понятий с наибольшими значениями показаны в табл.2.. Таблица 2 – Частотные характеристики. №. Слово (концепт). Количество (частота встречаемости), f. % в ядре (относительная частота), α. % в тексте (относительная частота), β. 1. 2. 3. 4. 5. Текст N 1 – t 1. 1. Комиссия. 3. 0,053. 0,041. 2. Фольга. 3. 0,053. 0,041. 3. США. 3. 0,053. 0,041. 4. Китай. 3. 0,053. 0,041. 5. Пошлина. 3. 0,053. 0,041. Текст N 2 – t 2. 1. Фольга. 7. 0,104. 0,071. 2. Пошлина. 5. 0,075. 0,051. 3. США. 4. 0,060. 0,040. 4. Год. 4. 0,060. 0,040. 5. Китай. 4. 0,060. 0,040. Текст N 3 – t 3. 1. Сталлоне. 5. 0,100. 0,067. 2. Рокки. 4. 0,080. 0,053. 3. Доска. 3. 0,060. 0,040. 4. Монумент. 2. 0,040. 0,027. 5. Мэр. 2. 0,040. 0,027. . На втором этапе вначале находятся меры семантической связности двух концептов с использованием WikiSim on Wikipedia . Для этого выполняется последовательный анализ пар статей: (, ), (, ), (, ). При проведении эксперимента анализ пары статей выполнялся на множествах пар понятий, перечисленных в табл. 2:. . В данной статье не описывается процесс построения пространственного графа, т.к. принцип его описание построения примерно аналогичен описанию, приведенному в работе [2]. Графы строятся для каждого концепта и . Анализ графов понятий и позволяет сформировать множество ссылок и . По формуле (1) вычисляются меры семантической связности концептов .. Выполнив анализ множества пар понятий по методу WikiSim on Wikipedia переходим к рассмотрению этого множества по методу WikiSim on Wiktionary . Принцип вычисления меры сходства тот же, что и в предыдущем случае. Результаты показаны в табл. 3.. Таблица 3 – Меры семантической связности концептов первых двух текстов. №. Понятия. Понятия. WikiSim on Wikipedia. . WikiSim on Wiktionary. . 1-й текст. 2-текст. 1. 2. 3. 4. 5. 1. Комиссия. Фольга. 0,03. 0,11. 2. Комиссия. Пошлина. 0,03. 0,04. 3. Комиссия. США. 0. 0,04. 4. Комиссия. Год. 0,03. 0,04. 5. Комиссия. Китай. 0. 0,02. 6. Фольга. Фольга. 1. 1. 7. Фольга. Пошлина. 0,04. 0,04. 8. Фольга. США. 0,04. 0,08. 9. Фольга. Год. 0. 0,02. 10. Фольга. Китай. 0,04. 0,02. 11. США. Фольга. 0,04. 0,08. 12. США. Пошлина. 0,03. 0,01. 13. США. США. 1. 1. 14. США. Год. 0,5. 0,05. 15. США. Китай. 0,4. 0,05. 16. Китай. Фольга. 0,04. 0,02. 17. Китай. Пошлина. 0,04. 0. 18. Китай. США. 0,4. 0,05. 19. Китай. Год. 0,08. 0,03. 20. Китай. Китай. 1. 1. 21. Пошлина. Фольга. 0,04. 0,04. 22. Пошлина. Пошлина. 1. 1. 23. Пошлина. США. 0,03. 0,01. 24. Пошлина. Год. 0,003. 0. 25. Пошлина. Китай. 0,13. 0. . Далее вычисляется гибридная мера семантической связности концептов по формуле (2):. . где i , j – индексы понятий первого t 1 и второго t 2 текстов соответственно;. l – порядковый номер пары концептов ;. весовые коэффициенты и вычисляются по формулам. , ;. .. Результаты вычислений показаны в табл. 4.. Таблица 4 – Гибридная мера семантической связности концептов первого и второго текстов. N пары концептов, l. Сумма,. . . . 1. 2. 3. 4. 5. 1. 0. 0,000. 0,000. 0,000. 2. 0. 0,000. 0,000. 0,000. 3. 0,04. 0,000. 1,000. 0,040. 4. 0,01. 0,000. 1,000. 0,010. 5. 0,01. 0,000. 1,000. 0,010. 6. 0. 0,000. 0,000. 0,000. 7. 0. 0,000. 0,000. 0,000. 8. 0,05. 0,000. 1,000. 0,050. 9. 0,01. 0,000. 1,000. 0,010. 10. 0,03. 0,000. 1,000. 0,030. 11. 0,1. 1,000. 0,000. 0,100. 12. 0,05. 1,000. 0,000. 0,050. 13. 0,02. 0,473. 0,526. 0,009. 14. 0,02. 1,000. 0,000. 0,020. 15. 0,08. 0,750. 0,250. 0,050. 16. 0,01. 1,000. 0,000. 0,010. 17. 0,01. 1,000. 0,000. 0,010. 18. 0,01. 0,000. 1,000. 0,010. 19. 0,09. 0,888. 0,111. 0,072. 20. 0,04. 0,500. 0,500. 0,020. 21. 0. 0,000. 0,000. 0,000. 22. 0. 0,000. 0,000. 0,000. 23. 0,06. 0,166. 0,833. 0,043. 24. 0,01. 1,000. 0,000. 0,010. 25. 0,04. 0,250. 0,750. 0,025. . Аналогичным образом находятся меры семантической связности концептов для пар текстов (t 1, t 3), (t 2, t 3).. На третьем этапе находится интегральная оценка семантической связности пары текстов или установление меры семантической ассоциации двух текстов. Мера семантической связности двух текстов рассчитывается по формуле:. . где wl – весовой коэффициент, позволяющий учесть частоту встречаемости понятий в ядре текста.. Данный коэффициент вычисляется последовательно следующим образом:. ; ,. ; ,. ,. где и относительные частоты встречаемости понятий первого и второго текста соответственно (см. табл. 2, графа 4).. и – нормализованные относительные частоты;. и – весовые коэффициенты, позволяющие учесть частоту встречаемости понятий первого и второго текстов соответственно.. Результаты вычисления меры семантической связности двух текстов показаны в табл. 5.. Таблица 5 – Результаты расчета меры семантической связности двух текстов ,. N пары концептов, l. . . Сумма. . . . . 1. 2. 3. 4. 5. 6. 7. 8. 1. 0,2. 0,29. 0,49. 0,4082. 0,5918. 0,1266. 0,0117. 2. 0,2. 0,209. 0,409. 0,489. 0,511. 0,1023. 0,0036. 3. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0037. 4. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0032. 5. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0018. 6. 0,2. 0,29. 0,49. 0,4082. 0,5918. 0,1266. 0,1266. 7. 0,2. 0,209. 0,409. 0,489. 0,511. 0,1023. 0,0041. 8. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0061. 9. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0018. 10. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0031. 11. 0,2. 0,29. 0,49. 0,4082. 0,5918. 0,1266. 0,0084. 12. 0,2. 0,209. 0,409. 0,489. 0,511. 0,1023. 0,0026. 13. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0925. 14. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0425. 15. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0333. 16. 0,2. 0,29. 0,49. 0,4082. 0,5918. 0,1266. 0,0042. 17. 0,2. 0,209. 0,409. 0,489. 0,511. 0,1023. 0,0041. 18. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0333. 19. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0061. 20. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0925. 21. 0,2. 0,29. 0,49. 0,4082. 0,5918. 0,1266. 0,0051. 22. 0,2. 0,209. 0,409. 0,489. 0,511. 0,1023. 0,1023. 23. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0023. 24. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0003. 25. 0,2. 0,167. 0,367. 0,545. 0,455. 0,0925. 0,0120. . 0,6070. . Аналогичным образом были рассчитаны меры семантической связности оставшихся двух пар текстов, результаты расчетов представлены в табл. 6.. Таблица 6 – Мера семантической связности трех пар текстов. Комбинация текстов. Мера семантической связности. 1 и 2 (, ). 0,607. 1 и 3 (, ). 0,054. 2 и 3 (, ). 0,05. . В таблице 1 приведены ссылки на использованные тексты. Анализ текстов показывает, чо полученные меры семантической связности текстов вполне адекватно отражают реальную картину.. Выводы. Ядром алгоритма вычисления меры семантической связности двух текстов является гибридная мера семантической связности двух понятий, основанная на применении метода WikiSim. Результаты исследования предложенного алгоритма показали, что его применение дает более точные результаты, чем применение классического WikiSim. В данной статье этот факт подтверждают таблицы 3-5.. В настоящей работе мы ограничились мощностью множества концептов текста равной пяти. Так как все тексты небольшого объема, то остальные концепты встречались не более одного раза. На наш взгляд, в дальнейшем необходимо выяснить зависимость мощности множества концептов от объема текста.. В статье не разделялись вики проекты на англоязычные и русскоязычные так как, если слова нет в русскоязычном Вики-проекте можно использовать его англоязычный аналог.. Совмещение формальных (структурированных – Викисловарь) и неформальных (неструктурированных - Википедия) источников знаний может привести к качественной оценке семантической связности между текстовыми единицами.. Благодаря быстрому и общедоступному развитию вики проектов, созданию алгоритма вычисления меры семантической связности текстов мы можем получать данные, соответствующие современным и актуальным на сегодняшний день реалиям.