У статті розглядається метод виявлення стійких словесних ланцюжків, особливості його роботи в різних модифікаціях і з різними параметрами. Запропонований метод можна впровадити як елемент автоматизованого робочого місця лінгвіста для складання словників словосполучень

Ключові слова: коллокація, околиця ланцюжка, частота зустрічальності

В статье рассматривается метод выявления устойчивых словесных цепочек, особенности его работы в различных модификациях и с различными параметрами. Предложенный метод можно внедрить как элемент автоматизированного рабочего места лингвиста для составления словарей словосочетаний

Ключевые слова: коллокация, окрестность цепочки, частота встречаемости

This article discusses a method to identify stable word chains, particularly his work in various versions and with different parameters. The proposed method can be implemented as part of a workstation for the linguist compiling dictionaries phrases

Keywords: collocation, neighborhood of the chain, frequency of occurrence

УДК 519.2: 81

ОБ ОСОБЕННОСТЯХ ПРИМЕНЕНИЯ СТАТИСТИЧЕСКИХ АЛГОРИТМОВ ВЫЯВЛЕНИЯ УСТОЙЧИВЫХ СЛОВЕСНЫХ ЦЕПОЧЕК

А В. Червяк*
Контактный тел.: 050-992-34-98
E-mail: alexche8@gmail.com

А.В. Вечур
Кандидат технических наук, доцент*
Контактный тел.: 050-514-33-17
E-mail: vechur@yahoo.com

Е.Л. Шевченко
Кандидат технических наук, доцент*
Контактный тел.: 050-574-83-72
E-mail: olena.l.shevchenko@gmail.com

В.Н. Ляпота
Ассистент*
Контактный тел.: 050-497-25-27
E-mail: vitaliy.lyapota@gmail.com
*Кафедра программного обеспечения ЭВМ
Харьковский национальный университет радиоэлектроники
пр. Ленина 14, г. Харьков, Украина

1. Введение

В настоящее время очень часто возникает проблема создания различных словарей словосочетаний. Такие словари необходимы для обучения русскому языку иностранцев [1], для автоматизированного перевода [2], разрешения лексической омонимии в алгоритмах автоматического анализа текстов [3], выявление паронимических ошибок [4] и т.д. Создание таких словарей – весьма ресурсоемкий процесс, требующий объединения усилий многих квалифицированных специалистов и большого количества времени. Таким образом любая, даже частичная, автоматизация данного процесса представляет практический интерес в решении этой задачи.

В этой работе были рассмотрены особенности применения статистического метода выявления устойчивых словесных цепочек, предложенного В.Д. Гусевым и Н.В. Саломатиной [5]. Метод имеет несколько вариаций и конфигурационных параметров, от выбора значения которых зависит качество его работы. Использование полученных в исследовании результатов позволит определить, на что влияют различные модификации метода и значения его конфигурационных параметров, и какая их комбинация является наиболее оптимальной в определенных ситуациях.

2. Анализ последних исследований и публикаций

2.1. Определение терминологии

В области алгоритмов выявления устойчивых словесных цепочек различные авторы и источники используют разрозненную терминологию и её ин-

терпретацию: словесная цепочка, словосочетание, коллокация. С точки зрения лингвистов различаются свободные и устойчивые словосочетания.

В наиболее широком понимании «словесная цепочка» – это последовательность из двух и более слов, связанных на основе некоторого множества критериев. Например, если таким критерием является грамматическая и смысловая согласованность, то словесная цепочка становится словосочетанием.

С лингвистической точки зрения словосочетание – это соединение двух или нескольких знаменных слов, связанных по смыслу и грамматически, служащее для расчлененного обозначения единого понятия [6]. В свободном словосочетании его элементы могут быть заменены их синонимами. Одно из слов в такой интерпретации является доминантой, а второе выбирается на основе первого для передачи смысла всего выражения [7]. В таком понимании словосочетания ключевую роль играет понятие «грамматической сочетаемости».

В отличие от первого понятия более узким в лингвистике является понятие «устойчивого словосочетания» (идиомы, фразеологизма) – это лексически неделимое, устойчивое в своем составе и структуре, целостное по значению словосочетание, воспроизводимое в виде готовой речевой единицы [8]. Его смысл фразеологизма не выводится из смыслов его компонент. Замена компонент в таком словосочетании не допускается, поэтому такое словосочетание является несвободным. В таком понимании словосочетания ключевую роль играет понятие «лингвистической устойчивости».

Наиболее общим и упрощенным является определение, использованное в работе В.П. Захаровым и М.В. Хохловой [8]: устойчивая цепочка слов – это комбинация двух или более слов, имеющих тенденцию к совместной встречаемости. Такие цепочки были названы коллокациями. Этот термин впервые введен в Словаре лингвистических терминов О.С. Ахмановой [9]. Такое определение является наиболее полезным и конструктивным для статистических методов выделения словосочетаний. Это так, потому что в нем критерии грамматической сочетаемости и лингвистической устойчивости могут быть заменены статистическими, измеряемыми статистическими метриками, например, частотой встречаемости. Поэтому коллокация – это словосочетание, устойчивое в статистическом смысле. Статистически устойчивым может оказаться как свободное словосочетание, так и фразеологизм. Изложенные в данной статье результаты исследования позволяют косвенно выяснить, в какой степени множество коллокаций соответствует множеству словосочетаний в лингвистическом смысле и какие их разновидности могут быть найдены с помощью этой методики.

2.2. Суть статистических методов выделения коллокаций

Общим в различных методах статистических стратегий выделения коллокаций является то, что они ищут некоторые цепочки слов, частота вхождения в текст которых должна превосходить частоту вхождения вариаций этих цепочек.

В своей статье [10] П.И. Браславский исследует методы нахождения «терминов-кандидатов» на словосочетания с помощью алгоритмов MaxLen, C-value, k-factor и пр., в которых основной упор делается на различные статистические критерии устойчивости.

В работе В.Д. Гусева и Н.В. Саломатиной [5] простая метрика «частота встречаемости» кандидата на словосочетание исследуется в сравнении с цепочкой слов, продолженной вправо или влево еще одним любым словом. Для более жесткого критерия отбора словесных цепочек было введено понятие «комбинаторная вариантность». То есть отбор теперь включает в себя не только анализ право- и левосторонних расширений, но так же цепочек слов которые получились в результате удаление, замены или добавления одного слова. Результаты такого подхода, по мнению авторов статьи [5], дают наиболее чистые с точки зрения лингвистики результаты, поэтому именно этот алгоритм взят за основу для проведения исследования.

3. Принятое определение устойчивости цепочки слов

Так как основная цель статьи – это исследование статистического метода выявления словосочетаний, то под этим термином далее подразумевается статистически устойчивое словосочетание (коллокация), а именно, словесная цепочка, длинной более или равной 2, не прерываемая знаками препинания, которая имеет большее число вхождений по сравнению с ее различными вариациями. А именно, с цепочками, в которых будет удалено, заменено или добавлено одно слово.

4. Примененный алгоритм выделения коллокаций

Данный алгоритм можно разбить на несколько этапов: предварительная обработка текста, выявление и отбор цепочек заданной длины, проверка цепочки на устойчивость. На рис. 1 изображена диаграмма деятельности выполнения алгоритма:

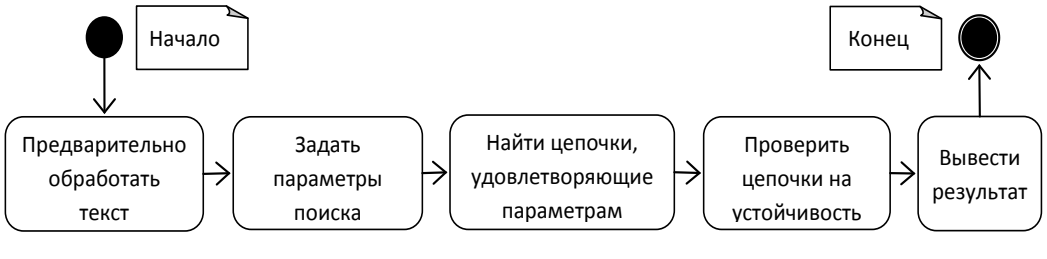

Рис.1. Общий вид алгоритма выделения коллокаций

4.1. Предварительная обработка текста

Перед выделением коллокаций текст должен пройти предварительную обработку.

Первоначальная обработка текста заключается в разбиении текста на блоки, не разделенные знаками препинания. Под знаками препинания подразумеваются [( ) : ; - ! ? . , \ / “”]. По умолчанию дефис считается обычной буквой алфавита. С учетом того, что слова, которые стоят через дефис, тоже могут образовывать цепочки слов, был проведен тест (см. п.5.2), в котором все дефисы были предварительно удалены. Также было предположено, что цифры, иностранные слова, типографические символы и т.д. не могут находиться в составе цепочки слов. Поэтому они также предварительно удаляются из текста.

Затем все слова в тексте приводятся к нормальной форме, что позволяет не учитывать различия в написания одного и того же слова в различном падеже, числе, времени и т.д. Для этой операции используется утилита mystem [11]. К сожалению, для многих слов однозначно определить нормальную форму, учитывая лишь морфологические признаки, невозможно.

Пример обработки словосочетания “В данной форме” приведен в табл. 1:

Таблица 1

Нормальные формы компонент словосочетания

В	в (предлог)
данной	давать (глагол) данный (наречие) Данна (имя собстенное)
форме	форма (существительное)

Правильной нормальной формой для слова «данной» является данный (наречие). Такой вывод можно сделать только при анализе всего словосочетания. Но так как русский язык не имеет четкого порядка слов в предложении, а количество частей речи, которые не могут стоять рядом (не комбинируются) достаточно ограничено, то нахождение нормальной формы слова машинным анализом, сканируя всё предложение, весьма ресурсоемко и затруднительно.

Для решения этой проблемы просто составляются и оцениваются все возможные варианты исследуемой цепочки. Таким образом, из искомого словосочетания были получены такие варианты цепочек: в давать форма, в данный форма, в Данна форма. Количество таких цепочек будет равно: $N=n_1*n_2*n_3...*n_n$, где n – количество возможных начальных форм i-го слова. Такой подход обусловлен тем, что одно и тоже сканируемое слово с разной лексической характеристикой на входе на выходе может иметь разное количество нормальных форм. При определении, что последовательность слов, составленная из таких неопределенных начальных форм, является коллокацией, эксперту будут выданы все её вариации.

4.2. Определение окрестности цепочки

После того, как были составлены все возможные комбинации цепочек с учетом неопределенности начальной формы слова, подсчитывается количество каждой такой цепочки в тексте. Так как устойчивыми могут быть только те цепочки, количество которых 2 или более, цепочки, встретившиеся единожды, отсеиваются уже на этом этапе, и дальнейшие операции будут проводиться только над теми, которые прошли этот отбор.

Проверка на устойчивость цепочки включает в себя нахождение её окрестности [5]. Окрестность цепочки – это вариации данной цепочки с учетом операций удаления, замены, добавления в нее одного слова. Например, окрестность цепочки *abc* приведена в табл. 2:

Таблица 2

Подвиды окрестностей цепочки

Удаление	ab, bc, ac
Замена	xbc, axc, xbc
Двустороннее и внутреннее расширение	xabc, axbc, abxc, abcx

Заметим, что операции замены и удаления не применимы к цепочкам длины 2.

Критерий устойчивости цепочки заимствован из работы [5]. Для каждого вида окрестности вычисляется отношение устойчивости:

$$\frac{\text{vic}}{\text{col}} \le \text{P},$$

где vic – количество вхождений цепочек окрестности, col – количество вхождений исходной цепочки, P – порог. Порог – эта величина, которая определяет жесткость отбора цепочек. Чем меньше порог, тем точнее будут результаты.

Если проверка хотя бы одного вида окрестности не проходит отбора – цепочка отсеивается. Если не была найдена ни одна цепочка определенной окрестности, то критерий отбора по этому виду окрестности считается выполненным. Если цепочка прошла все три проверки, либо для нее не нашлось окрестности, то она считается устойчивой.

5. Анализ экспериментальных результатов

Данным алгоритмом было обработано два текста: “Метод вызванных потенциалов мозга в американской психолингвистике и его использование при решении проблемы порядка слов в русском языке“ (12295 словоформ) из материалов конференции “Диалог 2006” и “Винни-Пух” Алана Милна в переводе Бориса Заходера (39806 словоформ). Для более подробного изучения особенностей работы алгоритмов был проведен ряд тестов. Для всех тестов, где иное не указано явно, порог P равен 0.5. Такое значение является оптимальным, потому что в этом случае цепочка считается устойчивой,

если элементы из её окрестности должны встречаться минимум в два раза меньше, чем сама цепочка.

5.1. Влияние различных составляющих окрестности цепочки на результаты работы алгоритма

В левой колонке указан выбранный вид окрестности. Анализируется изменение количества найденных коллокаций определенной длинны при каждом варианте окрестности (табл. 3).

Таблица 3

Результаты выделения коллокаций в тексте материалов «Диалог 2006»

Вид окрестности	Длина найденных цепочек			
	2	3	4	5
по 2х сторон. расширению	585	161	35	13
по 2х сторон. расширению + внутреннее расширение	569	153	35	13
по удалению	не применим	114	84	47
по замене	не применим	432	185	90
по полной окрестности	569	48	25	9

Исходя из результатов данного теста, можно увидеть, что количество коллокаций уменьшается с ростом количества слов, содержащихся в них, а разница уменьшения зависит от метода отбора. Если сравнивать результаты отбора при «двухстороннем расширении» и «двухстороннем + внутреннем расширении», то можно увидеть что они не намного отличаются. Разница в том, что в славянских языках словосочетания часто содержат в середине уточняющие слова, например, для двух слов: «тот же самый» (уточняет «тот самый»), «два референциальный электрод» (уточняет «два электрод»); для трех: «обращать особый внимание на» (уточняет «обращать внимание на»), для четырех и пяти слов таких цепочек не наблюдается.

При включении в окрестность цепочек, полученных удалением, не проходят цепочки такого типа: «метод вызывать потенциал» входит в «вызывать потенциал», «оборудование для» входит в «оборудование для запись» и т. д.

При поиске с учетом полной окрестности цепочки должны пройти отбор удалением, заменой и расширением. Например, цепочка «изучение весь этап» прошла отбор, потому что никаких ее вариаций в данном тексте нет, а цепочка «нагрузка на рабочий» не прошла - потому что хоть она и проходит отбор делецией (есть цепочка «нагрузка на», но она встречается меньше раз) , но есть расширенная цепочка «нагрузка на рабочий память», которая встречается чаще.

Стоит отметить, что наибольшую роль в отборе цепочек играет 2-х стороннее расширение, кроме цепочек длиной 3, где основную роль сыграло удаление. Это связано с тем, что большинство коллокаций длинной 3 являются различными уточнениями оттенков смысла одних и тех же коллокаций длиной 2.

Для текста «Винни-Пух» принцип изменения количества такой же.

Выводы. Среди цепочек длиной 2 и 3 часто попадаются свободные и требующие продолжения словосочетания. Цепочки длиной 4 и 5 в основном являются законченными свободными словосочетаниями, часто – специфической терминологией некоторой предметной области: «лабораторный оборудование для запись ВП».

Таким образом, самым оптимальным методом поиска является учет полной окрестности, при котором отбор наиболее жесткий и результат получается самым точным. При использовании этого метода стиль текста и его объем не имеют значения. Ослаблять жесткость отбора необходимо при желании получить «словосочетания, требующие продолжения» даже в узкоспециализированном тексте сравнительно небольшого объема. Наибольший вклад в отбор цепочки среди различных составляющих окрестности вносит 2х стороннее расширение.

5.2 Влияние различной интерпретации символа «дефис»

В данном тесте отсеивание коллокаций производилось по полной окрестности, но с двумя разными предобработками текста:

1. обычная (слова с дефисом учитываются как единая лексическая единица);
2. замена всех дефисов на пробелы (слова с дефисом распадаются на несколько лексических единиц).

Цель метода – определить наиболее оптимальное толкование символа «дефис»: как обычную значащую букву алфавита или как символ-разделитель самостоятельных слов.

При сравнении результатов этих двух подходов при втором появляются коллокации, которые целиком состояли из, или в состав которых входили сложносоставные слова или слова с частицами. Например: «какой то», «во первый», «куда то», «быть чем то занятый», «шестьдесят с что то дерево», «сверху вниз», «двухмодальный распределение негативность позитивность», «бум бум» и т. д. Как видно, такие цепочки не могут являться обычным словосочетанием, так как их написание раздельно является синтаксической ошибкой. Исчезают при втором варианте поиска, соответственно, цепочки со сложносоставными словами и словами с частицами. Например: «не так-то легко», «весь из-за то» и т. д. Новых цепочек появилось очень мало, и как правило они состоят из частей сложносоставных слов или частей слов с частицей. Например: в какой (входило в «какой-то»), ты где (входило в «ты где-то»), кристофер робин куда (входило «кристофер робин куда-то») и, следовательно, не представляют интереса как самостоятельная смысловая единица.

Выводы: Таким образом, интерпретация дефиса, как разделителя слов ведет только к увеличению «мусора» в результатах поиска коллокаций. Из этого следует, что применение данного подхода не является целесообразным.

5.3 Влияние учета/неучета служебных частей речи

Целью данного теста является выявление степени влияния служебных частей речи на виды словосочетаний, выделяемых алгоритмом в текстах. Поиск осуществляется в режиме отбора по полной окрестности (табл. 4, 5).

Таблица 4

Результаты выделения коллокаций в материалах «Диалог» 2006

Режим	Длина найденных цепочек			
	2	3	4	5
С учетом служебных частей речи	569	48	25	9
Без учета служебных частей речи	468	35	17	5

Комментарии. При замене служебных частей речи пробелами цепочек длинной 2 и 3 становится примерно в 1.2 раза меньше. Это свидетельствует о том, что 20% цепочек, которые находит алгоритм, содержат в себе служебные части речи, например: «мозг в», «в американский», «предложение с», «по правда говорить», «взад и вперед», «вытаскивать из голова». Результаты также свидетельствуют о том, что с ростом длины, процент цепочек со служебными частями речи увеличивается.

Среди полученных после удаления служебных частей речи новых цепочек есть верные и неверные по смыслу.

Правильными можно считать цепочки, в которых были удалены союзы, например, «маленький грустный кролик», полученную из «маленький и грустный кролик», «восприятие понимание» из «восприятие и понимание».

Цепочки, появившиеся в результате удаления предлогов, потеряли смысл, например, «общий знание мир», полученная из «общие знания о мире», «он пить кофе сахар», полученная из «он пить кофе с сахар», цепочка «оборудование запись», полученная из «оборудование для записи».

Выводы. Предварительно можно удалять только союзы. Предлоги удалять нельзя, т. к. они задают отношения между словами и часто входят в состав незаконченных словосочетаний.

5.4. Выделение общих коллокаций из нескольких текстов

Целью данного теста является проверка, на сколько могут пересекаться найденные в двух различных текстах словосочетания и в каких задачах данный тест может быть полезен.

В первой таблице содержаться результаты, полученные при анализе художественного и научного текстов, а во второй для двух художественных. Это сделано для того, чтобы увидеть, каким образом тематика текстов влияет на количество выявленных коллокаций.

Таблица 5

Результаты выделения общих коллокаций в текстах разного жанра

Текст	Длина найденных цепочек			
	2	3	4	5
«Диалог»	569	48	25	9
«Винни-Пух»	2832	125	97	28
Пересечение текстов	35	0	0	0

Комментарии. Общие коллокации были обнаружены только при длине цепочки 2. Были выявлены цепочка вида: «тот самый», «при тот», «который быть», «друг друг», «мы должный», т. е. в основном такие, в состав которых входят одно слово самостоятельной части речи (в основном местоимение) и одно слово служебной части речи.

Таблица 6

Результаты выделения общих коллокаций в текстах одного жанра

Текст	Длина найденных цепочек			
	2	3	4	5
«Алиса в стране чудес»	1105	50	12	0
«Винни-Пух»	2832	125	97	28
Пересечение текстов	284	0	0	0

Для текстов одного жанра ситуация схожа, но процент коллокаций длиной 2 увеличивается. Примеры совпавших цепочек: «это значит», «он становиться», «некоторый время», «я говорить», «очень обрадоваться», «становиться такой», «самый высокий» и т.д.

Выводы. Результаты проведенного теста подтверждают, что тест на пересечение может быть использован для:

а) выделения словосочетаний общего назначения, которые могут быть, например, присущи текстам разных жанров;

б) определения тематики текста на основе текста эталона;

в) для проверки алгоритма выделения словосочетаний. При этом общие словосочетания для двух и более текстов одинаковой или различной тематики с большей вероятностью могут быть найдены в уже существующих словарях, составленных лингвистами.

5.5. Сравнение результатов работы текста с готовым словарем словосочетаний.

Целью данного теста является проверка качества работы исследуемого в статье алгоритма в выбранной по результатам предыдущих тестов конфигурации. А именно, порог отбора P = 0.5, рассматриваются только

законченные словосочетания, дефис считается обычной буквой алфавита.

В качестве материала для проверки выбран текст конституции Российской федерации, а в качестве словаря – «Энциклопедический словарь конституционного права» (2000 терминов) [11].

Для оценки правильности найденных словосочетаний был использован критерий близости двух словосочетаний S_1 и S_2, заимствованный из работы Браславского [10]:

$$Sim(S_1,S_2) = |S_1 \cap S_2| / |S_1 \cup S_2|$$

Приведем пример вычисления степени близости некоторых пар словосочетаний:

Таблица 7

Пример вычисления степени близости пар словосочетаний

Кандидат на словосочетание	**Словосочетание из словаря**	**Sim**
Российская федерация	Российская федерация	1
Многонациональный народ	Власть народ	0.33
Федеральный собрание	Федеральный собрание российской федерации	0.5
Орган законодательный власти	Орган законодательной и судебной власти	0.6

По результатам проведенного эксперимента всего было найдено 132 кандидата на законченные словосочетания. Поиск в энциклопедическом словаре конституционного права (см. табл. 8) позволил с различной степенью близости подтвердить, что 55 из них являются словосочетаниями. Среди них степень близости 50% и более была выявлена у 47 цепочек.

Этот показатель может быть улучшен, если при подсчете степени близости не учитывать служебные части речи, тогда число совпавших словосочетаний увеличивается на 4.

Таблица 8

Оценка результатов эксперимента

Диапазон рассмотренной меры близости	Процент словосочетаний, подтвержденных по словарю	Из них по типам совпадения		Всего совпало со словарем (шт.)
		Полное совпадение (шт.)	Частичное совпадение (шт.)	
50-100%	36%	33	14	47 (из 132)

Рассмотрение 85 словосочетаний, не найденных в словаре, или со степенью близости меньше 50% показало, что они являются правильными, и не были найдены из-за неполноты словаря. Например, «экономическая деятельность», «частная жизнь», «родной язык», «медицинская помощь», «культурная ценность» и др. Ошибочными можно считать лишь две из 132 цепочек: «федерация предложение» и «предложение настоящий»

Вывод. По результатам теста можно сделать вывод, что алгоритм подтвердил свою адекватность и в выбранной конфигурации может быть успешно использован для выделения законченных словосочетаний из текста.

Выводы

Алгоритм выделения устойчивых словесных цепочек, предложенный В.Д. Гусевым и Н.В. Саломатиной [5], был проверен на реальных данных в нескольких конфигурациях. Результатом может быть вывод, что при всей своей простоте, качество его приемлемо для решения задачи выделения устойчивых словесных цепочек. Точность работы алгоритма на проанализированных текстах составляет 99%, что позволяет сделать вывод о перспективе выбранного статистического подхода. Отбор устойчивых цепочек слов необходимо проводить для полной окрестности, в которой для всех цепочек длиной 2, 4 и больше наибольший вес в отсеивание кандидатов вносит 2-х стороннее расширение и только для цепочек длиной 3 – удаление. Дефис необходимо считать обычным символом алфавита. Удаление союзов не сказывается на результатах алгоритма. Для поиска только законченных словосочетаний эффективно предварительно отсеивать последовательности, в которых служебные части речи стоят по краям цепочки-кандидата.

Литература

1. Учебный словарь сочетаемости слов русского языка [Текст] / под ред. П. Н. Денисова и В. В. Морковкина. — М.: Русский язык, 1978.
2. Коваль, С. А. Системы переводческой памяти и оценка их эффективности [Текст] / С. А. Коваль, О. Ф. Каткова // НТИ — 2002 — сер. 2, № 3 — С. 17–26.
3. Большаков, И. А. Какие словосочетания следует хранить в словарях? [Текст] / И. А. Большаков // Труды Межд. сем. Диалог'2002. — т. 2. — М.: Наука, 2002. — С. 61–69.
4. Гусев, В.Д. Анализ ошибок, не выявляемых автоматическими корректорами [Текст] / В. Д. Гусев, Н. В. Саломатина // тез. докл. II–й Межвузовской конференции «Квантитативная лингвистика и семантика» (КВАЛИСЕМ-99), Новосибирск, 12–15 октября, 1999. — С. 8–12.
5. Гусев, В.Д. Алгоритм выявления устойчивых словосочетаний с учетом их вариативности (морфологической и комбинаторной) [Текст] / В. Д. Гусев, Н. В. Саломатина // Труды междунар. конф. Диалог-2004, Верхневолжский, 2–7 июня 2004. — М.: Наука, 2004. — С. 530–535.
6. Справочник по русскому языку. Словарь лингвистических терминов [Текст] / Д. Э. Розенталь, М. А. Теленкова. — Харвест, 2008. — 624 с.

7. Хохлова, М.В. Экспериментальная проверка методов выделения коллокаций [Текст] / М. В. Хохлова // Инструментарий русистики: корпусные подходы. — Slavica Helsingiensia: 2008. — № 34. — С. 343–357.
8. Захаров, В. П. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке [Текст] / В.П. Захаров, М.В. Хохлова // Труды международной конференции «Диалог-2006». — 2006. — С. 137–143.
9. Ахманова, О. С. Словарь лингвистических терминов [Текст] / О. С. Ахманова. — 2-е изд. — М.: Советская энциклопедия, 1969 — 607 с.
10. Браславский, П. Сравнение пяти методов извлечения терминов произвольной длины [Текст] / П. Браславский, Е. Соколов. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008). — Вып. 7 (14). — М.: РГГУ, 2008. — С. 67–74.
11. О программе mystem [Электронный ресурс] / Режим доступа : \www/ URL: http://company.yandex.ru/technology/mystem/ — 10.06.2011 г. — Загл. с экрана.
12. Энциклопедический Словарь Конституционного Права [Текст] / под ред. Р. А. Мандрик — Новосибирск, 2010. — 666 с., 61145

В даній статті надана формальна модель семантичного пошуку в спеціалізованій електронній бібліотеці. Надана схема побудови онтології. Сформульовано леми для функцій інтерпретації термів і концепцій

Ключові слова: семантичний пошук, пошук інформації, онтології

В данной статье представлена формальная модель семантического поиска в специализированной электронной библиотеке. Представлена схема построения онтологии. Сформулированы леммы для функций интерпретации термов и концепций

Ключевые слова: семантический поиск, поиск информации, онтологии

This article presents a formal model of semantic search in a specialized electronic library. A scheme for constructing an ontology is presented. The lemmas for the functions of interpretation of terms and concepts are formed

Key words: semantic search, information retrieval, ontology

УДК 519.767.6

ФОРМАЛЬНАЯ МОДЕЛЬ СЕМАНТИЧЕСКОГО ПОИСКА В ЭЛЕКТРОННОЙ БИБЛИОТЕКЕ

З.В. Дударь
Кандидат технических наук, профессор, директор Центра
Центр последипломного образования*
Контактный тел.: (057) 702-18-05, 702-14-46
E-mail: fpo@kture.kharkov.ua

В.А. Белоконь
Аспирант**
Контактный тел.: (057) 702-18-05, 702-14-46
E-mail: fpo@kture.kharkov.ua

В.Г. Хильский
Магистрант
Контактный тел. (0625) 27-62-20, 063-243-84-33
E-mail: xv1975@mail.ru
**Кафедра программного обеспечения ЭВМ
*Харьковский национальный университет радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166

Введение

Развитие индустрии систем электронного документооборота, сопровождающееся ростом массивов обрабатываемых полнотекстовых документов, требует новых средств организации доступа к информации, многие из которых следует отнести к разряду систем искусственного интеллекта - систем обработки знаний. Основной задачей, возникающей при работе с полнотекстовыми базами данных, является задача поиска документов по их содержанию. Однако, ставшие традиционными средства контекстного поиска по вхождению слов в документ, представленные, в частности, поисковыми машинами в интернет, зачастую не обеспечивают адекватного выбора информации по запросу пользователя.

Первые информационно-поисковые системы (ИПС) появились более тридцати лет назад и с тех произошли существенные изменения, как в поисковых алгоритмах, так и в техническом оснащении. В настоящее время в поисковых системах используется релевантная модель оценки соответствия исследуе-