РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 685 044**(13) **C1**

(51) МПК
G06F 17/27 (2006.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(52) СПК
G06F 17/27 (2019.02)

(21) (22) Заявка: **2018124219, 03.07.2018**

(24) Дата начала отсчета срока действия патента:
03.07.2018

Дата регистрации:
16.04.2019

Приоритет(ы):
(22) Дата подачи заявки: **03.07.2018**

(45) Опубликовано: **16.04.2019** Бюл. № 11

Адрес для переписки:
117218, Москва, ул. Большая Черёмушкинская, 25, НИЦ "Курчатовский институт" - ИТЭФ, патентный отдел

(72) Автор(ы):
Доренская Елизавета Александровна (RU),
Семенов Юрий Алексеевич (RU)

(73) Патентообладатель(и):
Федеральное государственное бюджетное учреждение "Институт теоретической и экспериментальной физики имени А.И. Алиханова Национального исследовательского центра "Курчатовский институт" (НИЦ "Курчатовский институт"-ИТЭФ) (RU)

(56) Список документов, цитированных в отчете о поиске: **US 5369714 A, 29.11.1994. US 6332143 B1, 18.12.2001. US 6523026 B1, 18.02.2003. US 2014/0249804 A1, 04.09.2014. RU 2639684 C2, 21.12.2017.**

(54) СПОСОБ ОПРЕДЕЛЕНИЯ КОНТЕКСТА СЛОВА И ТЕКСТОВОГО ФАЙЛА

(57) Реферат:

Изобретение относится к области определения контекста слов и текстовых файлов. Технический результат заключается в повышении эффективности, достоверности и скорости определения контекста слова, текстового фрагмента и текстового файла. Технический результат достигается за счет подсчета расстояний между словами для определения контекстного значения слова с привлечением весовой функции и метрик слов по формуле $C_{k,n}=\sum_{i=1}^{m}(M_i \times f(L_i))$; где $C_{k,n}$ - мера, определяющая контекстное значение слова W1, индекс k для $C_{k,n}$ определяет, к какому из возможных значений W2 относится данная мера, где k=1,…,n, n - число возможных значений слова W1, где n=2÷3, Mi - метрика слова-характеристики W3, Li - расстояние от слова W2 до заданного слова W3, i - номер слова-характеристики в исследуемом тексте для слова W3, f(Li) - весовая функция от Li расстояний между словами W1, W2 и W3, m - число слов-характеристик, найденных в исследуемом текстовом файле. 2 н. и 5 з.п. ф-лы, 1 табл., 7 ил.

RU 2685044 C1

RU 2685044 C1

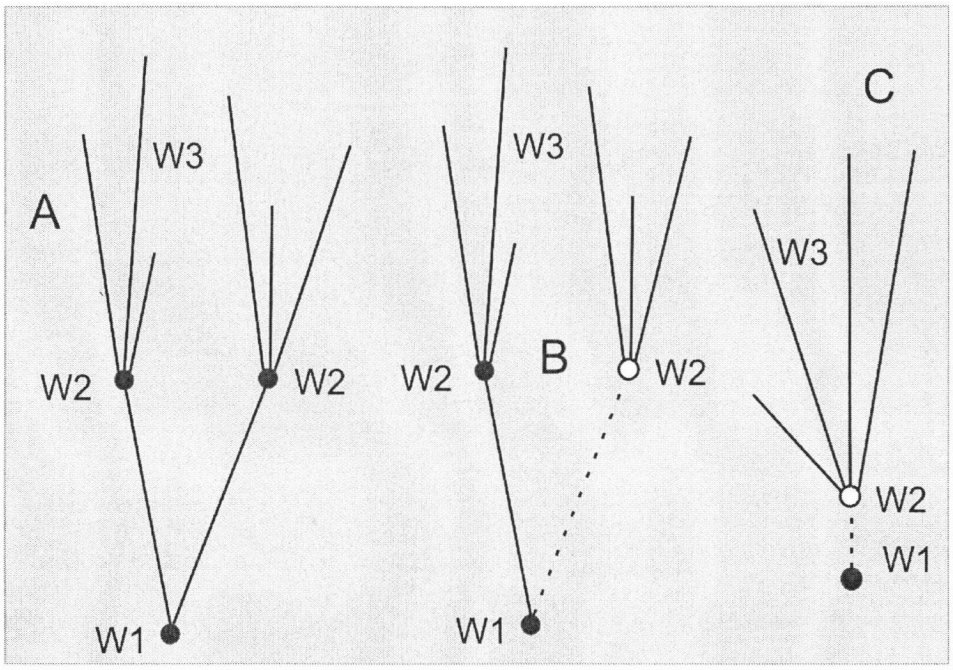

Рис. 1

RU 2685044 C1

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 685 044**(13) **C1**

(51) Int. Cl.
G06F 17/27 (2006.01)

(12) ABSTRACT OF INVENTION

(52) CPC
G06F 17/27 (2019.02)

(21) (22) Application: **2018124219, 03.07.2018**

(24) Effective date for property rights:
03.07.2018

Registration date:
16.04.2019

Priority:
(22) Date of filing: **03.07.2018**

(45) Date of publication: **16.04.2019** Bull. № **11**

Mail address:
117218, Moskva, ul. Bolshaya Cheremushkinskaya, 25, NITS "Kurchatovskij institut" - ITEF, patentnyj otdel

(72) Inventor(s):
Dorenskaya Elizaveta Aleksandrovna (RU),
Semenov Yurij Alekseevich (RU)

(73) Proprietor(s):
Federalnoe gosudarstvennoe byudzhetnoe uchrezhdenie "Institut teoreticheskoj i eksperimentalnoj fiziki imeni A.I. Alikhanova Natsionalnogo issledovatelskogo tsentra "Kurchatovskij institut" (NITS "Kurchatovskij institut"- ITEF) (RU)

(54) METHOD OF DETERMINING CONTEXT OF WORDS AND TEXT

(57) Abstract:

FIELD: data processing.

SUBSTANCE: invention relates to determining the context of words and text files. Technical result is achieved by counting distances between words to determine context value of word with involvement of weight function and metrics of words by formula $C_{k,n}=\Sigma^{m}_{i=1}(M_i \times f(L_i))$; where $C_{k,n}$ is a measure which determines context value of word W1, index k for $C_{k,n}$ determines to which of possible values W2 relates this measure, where k = 1,…,n, n – number of possible values of word W1, where n=2÷3, Mi is metric of word-characteristic W3, Li is distance from word W2 to given word W3, i is number of characteristic word in analyzed text for word W3, f(Li) is weight function of Li distances between words W1, W2 and W3, m is number of words-characteristics found in analyzed text file.

EFFECT: technical result is higher efficiency, reliability and speed of determining the context of a word, a text fragment and a text file.

7 cl, 1 tbl, 7 dwg

RU 2685044 C1

RU 2685044 C1

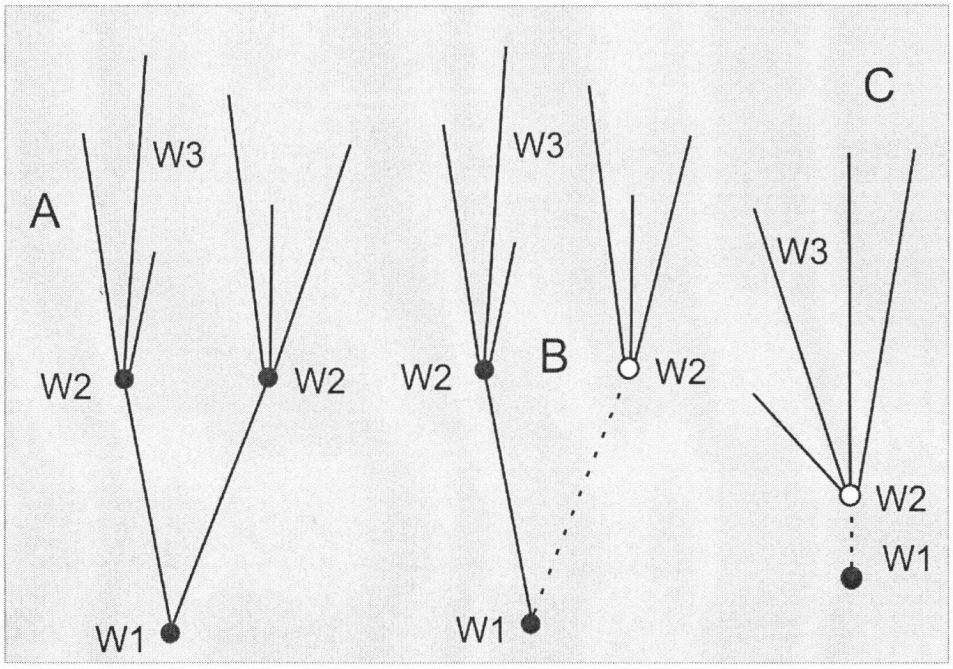

Рис. 1

RU 2685044 C1

RU 2685044 C1

Область техники

Данное изобретение относится к способам определения контекста слов и текстовых файлов и может быть использовано для анализа различных текстов на естественном языке. Данное изобретение имеет как минимум три области применения: поиск с целью анализа контекста запроса, в технологиях искусственного интеллекта, а также при создании и проверке алгоритмов путем машинного анализа текстовых описаний программ и инструкций, с целью интерпретации и исполнения машиной заданного условия. Такие описания могут сократить количество программных ошибок за счет того, что основной код будет генерироваться с помощью ЭВМ, согласно заданному текстовому описанию.

Уровень техники

Одно и то же слово может употребляться в разных значениях. Например, в русском языке слово «ключ» может иметь значения «ключ от замка», ключ как «родник», ключ как «разгадка» или криптоключ. Человек способен определять контекстное значение такого слова, анализируя соседние слова в предложении и сам текст в целом. Аналогичные пути используются в компьютерном анализе текстов и текстовых файлов. Ниже рассмотрены некоторые известные на данный момент методы.

Определение контекста - общего смысла слова и текстового файла на сегодняшний день очень важная задача, так как от интерпретации текстов машиной зависит ее способность работать с текстами. Эти проблемы сейчас относятся к Al-полным задачам, требующим сильного искусственного интеллекта. Ими сейчас занимается NLP (Natural Language Processing - Обработка естественного языка) - общее направление искусственного интеллекта и математической лингвистики, изучающее компьютерный анализ и синтез естественных языков. Для искусственного интеллекта под анализом подразумевается понимание и разбор семантического значения текстов на естественном языке, а под синтезом - их грамотная генерация. Повышение удобства взаимодействия компьютера и человека является главной задачей на сегодняшний день в данной области. [Могилев А.В. и др., Технологии обработки текстовой информации. Технологии обработки графической и мультимедийной информации СПб.: БХВ-Петербург, 2010, стр 175].

Распознавание контекста часто осуществляется за счет семантических сетей и онтологий. Семантическая сеть - это информационная модель предметной области, представляющая собой ориентированный граф, вершины которого - объекты предметной области (понятия, события, свойства, процессы), а ребра -связи между ними [Roussopoulos N.D. и др., A semantic network model of data bases // Department of Computer Science, University of Toronto, 1976, TR No 104].

Понятиями называются отраженные в мышлении самые важные свойства, связи и отношения предметов или явлений. Понятием также является мысль или система мыслей, которая выделяет и обобщает предметы определенной группы согласно одинаковым, а также и специфическим признакам, в отношении них [Большая советская энциклопедия. - М.: Советская энциклопедия 1969-1978, том 20, стр. 1047].

Изобретение, описанное в патентном документе RU 2632126, опубл. 02.10.2017, относится к средствам предоставления контекстуальной информации, относящейся к документу, с использованием семантической сети. Однако в данном документе не определено, как оценивается контекстуальная релевантность объекта и как выполняется контекстуальный поиск. Из приведенного описания указанного документа ясно, что в оценку контекста вовлечен человек (стр. 30), который передает серверу нужные данные.

В отличие от патентного документа RU 2632126, опубл. 02.10.2017, в предложенных нами способах, раскрытых в настоящей заявке, вся работа происходит на одном компьютере, где анализируется текст и определяется контекстное значение слов и текстовых файлов. В том числе, за счет формулирования числового способа оценки контекста, предложенная нами методика применима как к отдельным словам или текстовым фрагментам, так и к тексту в целом, что позволяет эффективно определить контекст, предоставляя возможность числовой оценки вероятности правильности выбора контекстного значения.

Наиболее близким решением является изобретение, описанное в патентном документе US 9,632,999 B2, опубл. 25.04.2017, который относится к определению контекста текстового фрагмента с помощью семантического анализатора и семантических сетей. Текст в нем анализируется слово-за-словом с привлечением технологии семантических сетей, где учитываются смысловые связи отдельных слов.

Основными недостатками указанного патентного документа US 9,632,999 B2, опубл. 25.04.2017, являются следующие:

- способ применим скорее к фрагментам текста, чем к отдельным словам;
- способ довольно трудоемок и по этой причине с помощью него сложно определять контекст больших фрагментов текста;
- определяется только приблизительный смысл слова.

Техническая проблема заключается в трудоемкости проведения семантического анализа и связана с тем, что обработка текста требует слишком больших временных и машинных ресурсов.

Предложенные нами способы вычисления контекстного значения облегчают и ускоряют контекстный анализ, а также снижают затраты машинных ресурсов, в целом повышая эффективность и достоверность анализа. Кроме того, эффективность определения контекста обеспечивается количественной оценкой вероятности правильности полученного результата.

Раскрытие изобретения

Изобретение относится к анализу контекстных значений слов и текстовых файлов путем определения, к какому смысловому значению относится слово, текстовый фрагмент и/или текстовый файл.

Технический результат заключается в повышении эффективности, достоверности и скорости определения контекста слова, текстового фрагмента и текстового файла. Эффективность заключается в увеличении объективности оценки контекста слова, фрагмента текста или текстового файла в целом за счет получения количественных значений контекстуальной близости слов, что не достижимо при осуществлении ранее известных способов.

Контекстное значение слова определяется с учетом анализа тематического словаря семантически связанных слов для разных контекстов. Возможно анализировать различные тексты с целью выявления контекстного смысла, в том числе использовать предложенные нами способы в поисковых системах с целью анализа контекста запроса пользователя и формирования соответствующей поисковой выдачи. Способы настоящего изобретения включают использование имеющегося тематического словаря из 3-х колонок и подсчет расстояний между словами для определения контекстного значения слова с привлечением весовой функции и метрик слов. Контекстное значение фрагмента текста или файла в целом определяется путем суммирования значений $C_{k,n}$ для каждого из найденных в нем слов из первой колонки таблицы и сравнения вычисленных мер между собой. Наибольшие из них и будут определять контекстное

значение фрагмента и/или текста в файле.

Задачей, на решение которой направлено данное изобретение, является упрощение процесса определения контекста слова естественного языка, встречающегося в тексте файла, и всего текстового файла в целом. При решении поставленной задачи был достигнут указанный выше технический результат.

Данная задача решается за счет использования специализированных графов или таблиц. Каждое слово W1 (см. таблицу 1) является начальной точкой графа, от которой идут ветвления в сторону его вершин (слов W2). Эти вершины раскрывают контекстный смысл слова W1. Слова W2 имеют семантические связи со словами-характеристиками уровня W3, благодаря которым, в основном, и определяются контекстные значения слов W1.

Рассмотрим цепочки семантических связей текстового файла (см. рис. 1). Слово W1 может иметь два или более значений, зависящих от контекста и определяемых словами W2. Слова W2 могут и отсутствовать в тексте файла, в этом случае учитываются расстояния от W1 до каждого слова из списка W3, сопряженного со смыслом отсутствующего слова W2. Можно предположить, что чем ближе слово-характеристика к слову вышестоящей вершины графа, тем с большей вероятностью оно определяет контекст слова-значения. Наличие слова из третьей колонки, размещенного в тексте ближе к слову из второй колонки, должно влиять на выбор контекстного значения слова сильнее, чем в случае слов, размещенных дальше.

Для определения контекста вычисляется $C_{k,n} = \sum_{i=1}^{m}(M_i \times f(L_i))$, где C - мера, определяющая контекстное значение слова W1, L - расстояние между словом, например, "компьютер" (W2) и "отладка" - W3 (см. табл.1), M_i - метрика слова-характеристики W3, L_i - расстояние от слова W2 до заданного слова W3, i - номер слова-характеристики в исследуемом тексте для слова W3, $f(L_i)$ - весовая функция от L_i, m - число слов-характеристик, найденных в исследуемом текстовом файле; n - число возможных значений слова W1 (чаще всего n=2÷3), n равно числу слов W2 семантического дерева для слова W1. $f(L_i)$ может быть в простейшем случае равно $1/L_i$, а для небольших документов $f(L_i)=1$ (см. рис. 2,А). L определяется числом слов N размещенных между словом W2 и одним из слов W3 (L=N+1). В отсутствии слова W2, расстояние L отсчитывается от позиции слова W1. Списки слов W3, соответствующие разным словам W2, могут перекрываться.

Индекс k для C определяет, к какому из возможных значений W2 относится данная мера (k=1, … n) (см. вторую колонку таблицы 1).

Значение слова W2, соответствующего слову W1, с большим значением C в контекстном смысле считается предпочтительным, именно оно определяет контекстное значение для W1. Таблица никак не связана с каким-либо конкретным текстом, контекстные значения слов которого исследуются. Полнота таблицы сильно влияет на точность определения контекста.

Если длина документа превышает одну страницу (~400 слов), оптимальной может оказаться весовая функция рис. 2,Б, которая для малых расстояний обеспечивает линейный вклад метрик, как на рис. 2,А, а при больших расстояниях приобретает вид 1/L.

Таблица 1. Фрагмент таблицы корневых слов (W1), слов-значений (W2) и слов-характеристик (W3)

Корневое слово W1	Слова-значения W2	Слова-характеристики (W3)	Метрика [M]
Программа	компьютер	программирование	70
		отладка	60
		тестирование	40
		подпрограмма/subroutine	30
		объект	15
		файл	26
		прерывание	40
		Оперативная память	70
		переменная	30
		константа	20
		SSD	30
		массив/array	50
		библиотека (программ)	15
		язык (программирования - название)	60
	обучение	пособие	45
		преподаватель	50
		учащийся	95
		учебник	90
		дистанционное	70

Таблица должна быть создана заранее и никак не должна зависеть от исследуемого текста. Содержимое таблицы должно храниться в банке данных, что облегчит доступ.

В таблицу заносятся только слова W1, имеющие два или более контекстных значений. Исключение может быть сделано для таблиц, предназначенных для распознавания контекста документов. Полная таблица даже для отдельной области знаний будет в сотни и тысячи раз больше.

Семантически связанные слова, вписываемые в таблицу и граф, и составляют семантическую сеть. Для расчета контекста используется программа для компьютера, которая берет данные из этой таблицы и согласно этим данным рассчитывает контекстные значения.

Один из предложенных в настоящей заявке способов предусматривает проведение на компьютерном устройстве количественной оценки контекстных значений отдельных слов в текстовом файле путем численного анализа семантического графа слов в документе, где способ включает:

i. предоставление текстового файла для анализа,

ii. использование имеющегося тематического словаря: в первой колонке -слова (W1), для которых определяется контекст, во второй - варианты возможного значения контекста (W2), в третьей - слова (W3), семантически связанные с W2,

iii. подсчет расстояний между словами для определения контекстного значения слова с привлечением весовой функции и метрик слов по формуле $C_{k,n} = \sum_{i=1}^{m}(M_i \times f(L_i))$; где

$C_{k,n}$ - мера, определяющая контекстное значение слова W1, индекс k для $C_{k,n}$ определяет, к какому из возможных значений W2 относится данная мера, где k=1, …, n, n - число возможных значений слова W1, где n=2÷3, M_i - метрика слова-характеристики W3, L_i - расстояние от слова W2 до заданного слова W3, i - номер слова-характеристики в исследуемом тексте для слова W3, $f(L_i)$ - весовая функция от U расстояний между словами W1, W2 и W3, m - число слов-характеристик, найденных в исследуемом текстовом файле;

iv. определение контекстного значения с учетом расстояния между корневым словом W1 и словами-значениями W2 или словами-характеристиками W3, расстояние L исчисляется количеством слов N, размещенных между корневым словом W2 и словом-характеристикой W3, L=N+1,

v. в случае если слово W3 встречается в текстовом файле несколько раз, вклады от каждого из этих слов войдут в указанную сумму, при этом списки слов W3, соответствующие разным словам W2, могут перекрываться, слово W2, для которого получена наибольшая сумма, и определяет контекстное значение слова W1.

Контекстное значение базового слова W1 определяют даже в отсутствии одного или даже всех слов-значений W2, за счет наличия в текстовом файле слов из набора W3, соответствующих отсутствующим словам W2.

Область определения контекстного значения задается с помощью весовой функции, путем конфигурирования программного обеспечения или непосредственно самой программой.

Для определения достоверности вычисленного контекстного значения слова может использоваться либо распределение плотности вероятности для C, либо неравенство Чебышева.

Другой из предложенных в настоящей заявке способов направлен на количественную оценку контекстных значений текстовых файлов или фрагментов текста из них и также проводится путем численного анализа семантического графа слов в документах, выполняемого на компьютерном устройстве. Способ предусматривает осуществление указанных стадий i-v вышеописанного способа и определение контекстного значения текстового файла или фрагмента его текста путем суммирования значений $C_{k,n}$ для каждого из найденных в нем слов из первой колонки таблицы и сравнение вычисленных мер между собой, наибольшие из них и будут определять контекстное значение текстового файла или его текстового фрагмента.

Фрагментом текстового файла является неполная часть текста в файле, например, один или несколько абзацев или одна, или несколько страниц.

Для определения контекстного значения текстового файла вычисляется сумма величин $C_{k,n}$ для всех слов W2 и W3, где весовая функция расстояний между словами W1, W2 и W3. Слово W1, для которого получено наибольшее значение суммы C, и определяет контекст текстового файла или его фрагмента.

Краткое описание чертежей

На рис. 1 изображены варианты анализируемых семантических сетей, где W1 обозначает слово из первой колонки таблицы 1, контекст которого мы определяем, W2 - слово из второй колонки таблицы, которое представляет собой одно из нескольких возможных значений контекста для данного слова, W3 - слово из третьей колонки (всегда связано с каким-либо из слов W2).

Вариант А на рис. 1 предполагает наличие в тексте файла корневого слова W1, которое может иметь разные контекстные значения, определяемые словами W2. Слова-значения W2 могут в текстовом файле быть или отсутствовать (рис. 1,В).

Предполагается, что каждому из слов-значений W2 соответствует некоторое число слов-характеристик W3, именно они и определяют выбор контекстного значения слова W1 за счет выбора одного из слов-значений W2. Секция рис. 1,С иллюстрирует вариант оценки контекста текстового файла в отстутствии слова W2.

На рис. 2 (А, Б) приведены возможные формы весовых функций.

f(L) - весовая функция

L - расстояние от слова W2 до W3.

Весовая функция рис. 2,А относится к случаю коротких текстов, а рис. 2,Б - к длинным текстам (более одной страницы).

На рис. 3 изображен пример фрагмента семантической сети в виде графа, иллюстрирующего таблицу 1. На нем отображены слова из первой (W1), второй (W2) и третьей (W3) колонок данной таблицы. В самом его верху находится корневое слово W1 «программа». От него идут ветвления к словам W2, которые обозначают контекстный смысл слова W1 "программа". Метрика характеризует близость слова к контекстному значению с которым оно связано и обозначается буквой М. Слова во второй колонке всегда имеют метрику равную 100. Поэтому для слов «компьютер» и «обучение» М=100. Далее от них идут ветвления графа к словам из третьей колонки W3. Это все прочие слова в таблице. Под каждым из этих слов указана его метрика (<100) и они сильно влияют на контекст, определяемый для слова W1.

На рис. 3 отображен фрагмент схемы, поскольку контекстных смыслов у слова «программа» может быть много. Существует также много слов-характеристик, помимо указанных в колонке W3 таблицы 1. Графы такого вида и являются основой семантической сети, используемой нами для оценки контекстного значения слов и документов.

На рис. 4 приведена предпочтительная реализация работы программы, которая может быть создана с использованием предложенных способов.

1. Производится запуск программы;

2. Программа считывает содержимое текстового файла и заносит его в массив;

3. Массив разбивается на слова. Каждое слово текста - отдельный член массива;

4. Программа ищет нужное слово W1, если оно не найдено - прекращает свою работу;

5. Если хоть одно слово W1 найдено в тексте, программа ищет слова из колонок W2 и W3, во всех падежах и числах, чтобы проводить с ними нужные операции;

6. Программа вычисляет расстояние от каждого слова W1 до слов W2 и W3;

7. Программа присваивает метрику найденным словам из колонок W2 и W3 (у всех слов W2 метрика равна 100, для слов W3 задается уникальное значение в таблице);

8. Вычисляется величина С для каждого из заданных контекстных значений (слов W2);

9. Вычисленные контекстные значения программа сравнивает между собой (согласно наибольшему и определяется контекст документа);

10. Результат выводится пользователю.

Примеры осуществления способов согласно настоящему изобретению.

Предложенные способы осуществляются следующим образом. Контекст слова рассчитывают путем анализа слов из второй колонки W2 с помощью слов из третьей колонки W3. Для расчета используют метрику и расстояние от анализируемого слова из колонки W2 до слов, связанных с ним из третьей колонки W3, найденных в тексте. Если какое-то слово-значение W2 для базового слова W1 в тексте документа отсутствует, то для расчета расстояния L в качестве начала отсчета используется положение слова W1. Контекстное значение слова W1 определяют по наибольшему значению суммы С

для соответствующего слова W2.

Для больших документов контекст каждого конкретного слова W1 может оказаться разным для разных областей документа. Размер области может быть настраиваемым, с дискретом в одну страницу (~400 слов). При этом можно варьировать начало и размер области и отслеживать вариации значений С и контекстного значения конкретного слова W1.

Определяют контекст фрагмента текстового файла или файла в целом по наибольшему контекстному значению (С) для анализируемых слов первой колонки W1. Списки слов W1, W2, W3 формируются с учетом вариаций слов по числам и падежам.

Расчеты и выбор контекстного значения могут быть осуществлены с помощью специализированного ПО, например, авторской программы на языке Perl под ОС Linux, зарегистрированной в ФИПС под номером RU 2018615758 от 16.05.2018.

Реализация способов позволяет повысить эффективность, достоверность и скорость определения контекста слова, текстового фрагмента и текстового файла.

Испытание алгоритма на полутора десятках текстовых файлах показало, что достоверная оценка получается в более чем 95% случаев. Сравнение проводилось для результатов оценки программой и человеком, а также путем численной оценки вероятности полученного результата. В частности, для документа о протоколе "SET и другие системы осуществления платежей" с числом слов 40631 и числом W3=213 получено отношение вероятностей оценки контекста для выражений "компьютерная программа" и "программа реализации проекта" = 60.

(57) Формула изобретения

1. Способ количественной оценки контекстных значений отдельных слов в текстовом файле путем численного анализа семантического графа слов в документе, выполняемый на компьютерном устройстве, где способ включает:

предоставление текстового файла для анализа,

использование имеющегося тематического словаря: в первой колонке - слова (W1), для которых определяется контекст, во второй - варианты возможного значения контекста (W2), в третьей - слова (W3), семантически связанные с W2,

подсчет расстояний между словами для определения контекстного значения слова с привлечением весовой функции и метрик слов по формуле $C_{k,n}= \sum_{i=1}^{m}(M_i \times f(L_i))$; где $C_{k,n}$ - мера, определяющая контекстное значение слова W1, индекс k для $C_{k,n}$ определяет, к какому из возможных значений W2 относится данная мера, где k=1,…,n, n - число возможных значений слова W1, где n=2÷3, Mi - метрика слова-характеристики W3, Li - расстояние от слова W2 до заданного слова W3, i - номер слова-характеристики в исследуемом тексте для слова W3, f(Li) - весовая функция от Li расстояний между словами W1, W2 и W3, m - число слов-характеристик, найденных в исследуемом текстовом файле;

определение контекстного значения с учетом расстояния между корневым словом W1 и словами-значениями W2 или словами-характеристиками W3, расстояние L исчисляется количеством слов N, размещенных между корневым словом W2 и словом-характеристикой W3, L=N+1,

в случае если слово W3 встречается в текстовом файле несколько раз, вклады от каждого из этих слов войдут в указанную сумму, при этом списки слов W3, соответствующие разным словам W2, могут перекрываться, слово W2, для которого получена наибольшая сумма, и определяет контекстное значение слова W1.

2. Способ по п. 1, отличающийся тем, что контекстное значение базового слова W1

определяют в отсутствии одного или всех слов-значений W2 за счет наличия в текстовом файле слов из набора W3.

3. Способ по п. 1, отличающийся тем, что область определения контекстного значения задается с помощью весовой функции, путем конфигурирования программного обеспечения или непосредственно самой программой.

4. Способ по любому из пп. 1-3, отличающийся тем, что для определения достоверности вычисленного контекстного значения слова используют распределение плотности вероятности для C или неравенство Чебышева.

5. Способ количественной оценки контекстных значений текстовых файлов или их текстовых фрагментов путем численного анализа семантического графа слов в документах, выполняемый на компьютерном устройстве, где способ предусматривает осуществление способа по любому из пп. 1-4 и определение контекстного значения текстового файла или фрагмента его текста путем суммирования значений $C_{k,n}$ для каждого из найденных в нем слов из первой колонки таблицы и сравнения вычисленных мер между собой, наибольшие из них и будут определять контекстное значение текстового файла или его текстового фрагмента.

6. Способ по п. 5, отличающийся тем, что фрагментом текстового файла является неполная часть текста в файле: один или несколько абзацев, или одна, или несколько страниц.

7. Способ по любому из пп. 5-6, отличающийся тем, что для определения контекстного значения текстового файла или его фрагмента вычисляется сумма величин $C_{k,n}$ для всех слов W2 и W3, где весовая функция расстояний между словами W1, W2 и W3, и слово W1, для которого получено наибольшее значение суммы C, и определяет контекст текстового файла или его фрагмента.

1

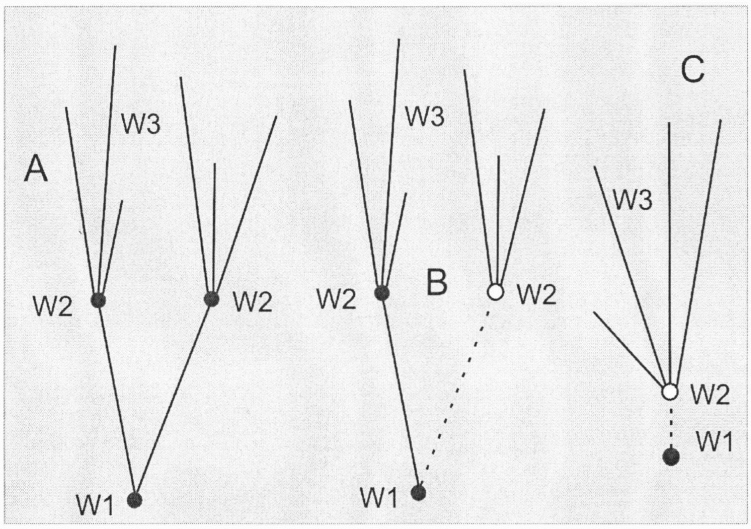

Рис. 1, A, B и C.

2

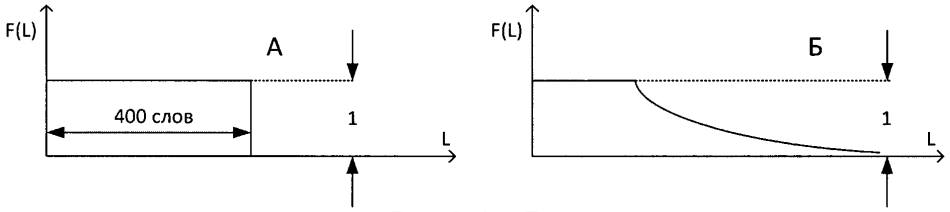

Рис. 2, А и Б.

3

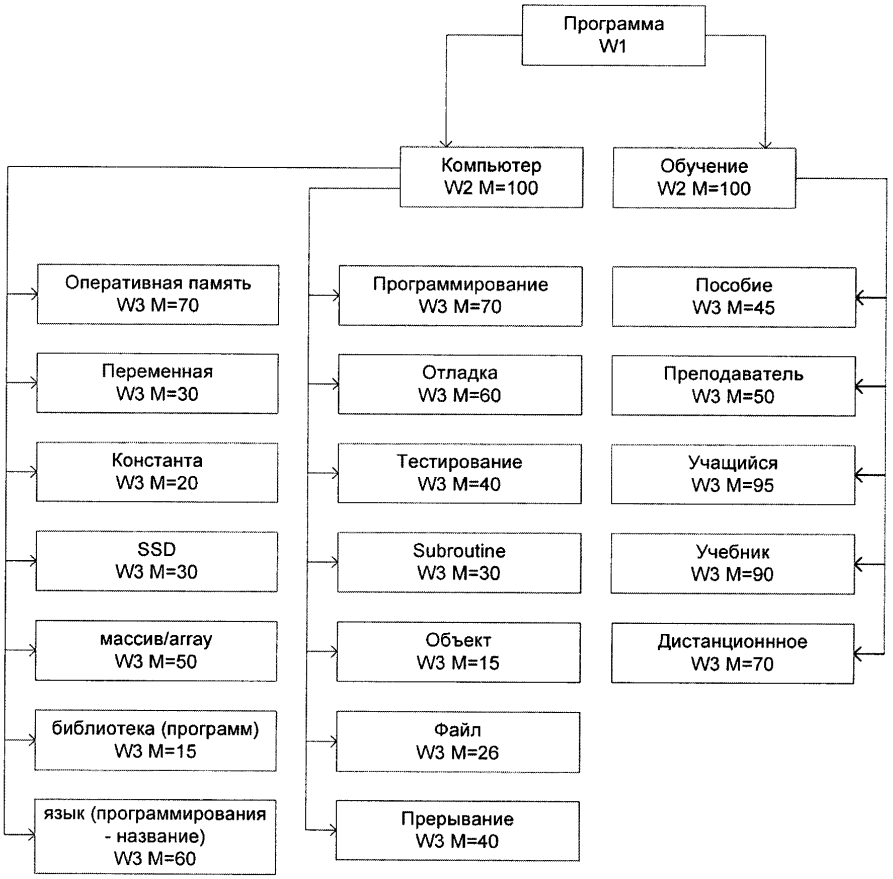
Программа
W1
Компьютер
W2 M=100
Обучение
W2 M=100
Оперативная память
W3 M=70
Переменная
W3 M=30
Константа
W3 M=20
SSD
W3 M=30
массив/array
W3 M=50
библиотека (программ)
W3 M=15
язык (программирования
- название)
W3 M=60
Программирование
W3 M=70
Отладка
W3 M=60
Тестирование
W3 M=40
Subroutine
W3 M=30
Объект
W3 M=15
Файл
W3 M=26
Прерывание
W3 M=40
Пособие
W3 M=45
Преподаватель
W3 M=50
Учащийся
W3 M=95
Учебник
W3 M=90
Дистанционнное
W3 M=70

Рис. 3.

4

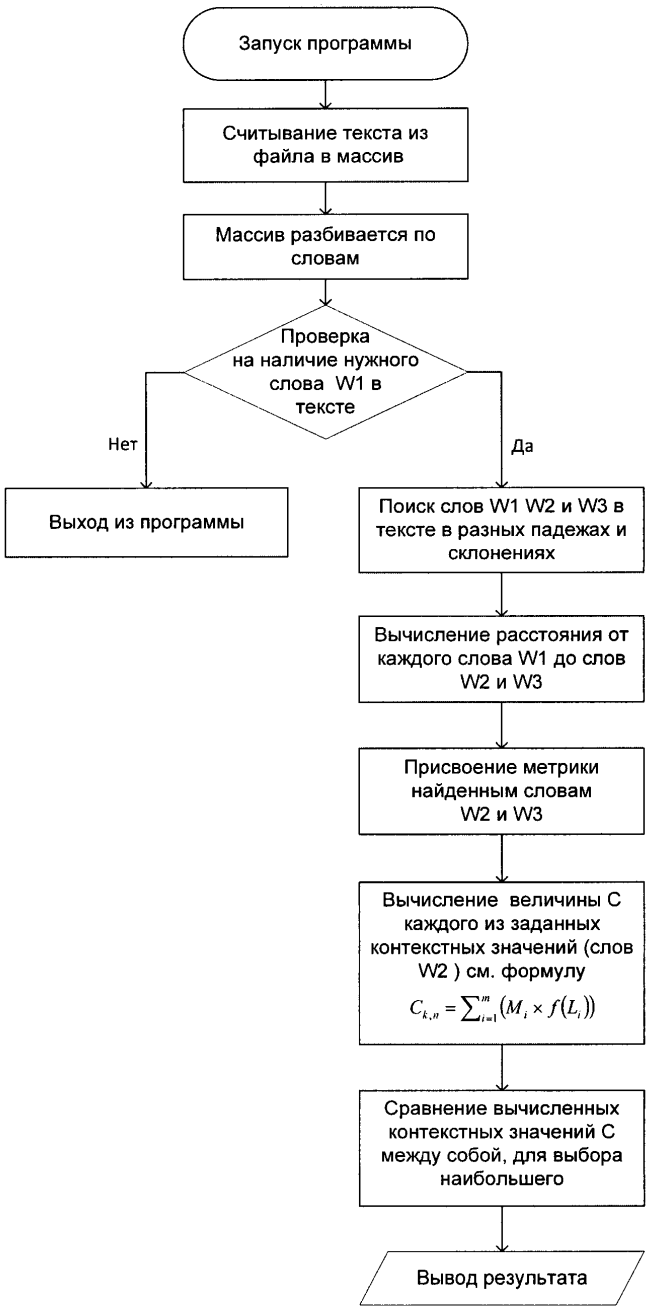
Запуск программы
Считывание текста из файла в массив
Массив разбивается по словам
Проверка на наличие нужного слова W1 в тексте
Нет
Да
Выход из программы
Поиск слов W1 W2 и W3 в тексте в разных падежах и склонениях
Вычисление расстояния от каждого слова W1 до слов W2 и W3
Присвоение метрики найденным словам W2 и W3
Вычисление величины C каждого из заданных контекстных значений (слов W2) см. формулу
$C_{k,n} = \sum_{i=1}^{m} (M_i \times f(L_i))$
Сравнение вычисленных контекстных значений C между собой, для выбора наибольшего
Вывод результата

Рис. 4.