3 БАЗЫ ЗНАНИЙ, ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ, ЭКСПЕРТНЫЕ СИСТЕМЫ, СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

Найханов Н.В., Дышенов Б.А.

ОПРЕДЕЛЕНИЕ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ ПОНЯТИЙ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ ССЫЛОК ВИКИПЕДИИ

Аннотация: *Предметом исследования является семантическая близость понятий. Объектом исследования меры семантической близости понятий. Авторы рассматривают такие аспекты темы как обоснование выбора фоновых знаний, построение ссылочного графа и измерение связанности между понятиями. В более ранних работах авторов семантическая близость вычислялась на основе статистических характеристик с применением различных методов контекстного анализа, например, латентно-семантического анализа. Данная работа является первым опытом работы со ссылочными методами определения семантической близости. Поэтому фокус сделан на простоту вычисления меры. В статье определение семантической близости основывается на методе WLM (Wikipedia Link-based Measure) и меры близости по отдельным типам ссылок М.И. Варламова, А.В. Коршунова. В отличие от известных мер семантической близости, основанных на использовании базы данных Википедии, предложенная в работе мера использует простые ссылки статей базы данных Википедии типа "См. также" (See also) и "Ссылки" (Links, External links). Такой подход позволяет повысить производительность алгоритма и применять в задачах, требующих не высокой точности результата, а большей производительности алгоритма. К таким задачам можно отнести установление соответствия между компетенциями образовательного стандарта и аннотациями дисциплин учебного плана или задачу анализа ответов студентов на открытые по форме вопросы. Разработанная мера является дешевой, достаточно точной и доступной.*

Ключевые слова: *понятие, семантическая близость понятий, фоновые знания, база данных Википедии, структура статьи Википедии, ссылка, ссылочный граф, расстояние между понятиями, индексация графа, м*ера,основанная на ссылках

Abstract: *The research question is the semantic relatedness of terms. The target of research is measure the semantic relatedness of terms. The authors consider such aspects as the rationale for the choice of the theme of background knowledge, the construction of a graph of links and measurement of relatedness between concepts. In earlier studies the authors of semantic proximity is calculated*

DOI: 10.7256/2305-6061.2016.3.19560

based on the statistical characteristics using different contextual analysis methods, such as latent semantic analysis. This work is the first experience with the reference methods for determining a semantic relatedness. Therefore, the focus placed on ease of calculation steps. Evaluation semantic similarity is based on the WLM method and proximity measure for separate types of references of M. I. Varlamov, A.V. Korshunov. In contrast to the well-known measures of semantic proximity, based on the use of Wikipedia proposed in the measure uses a simple links Wikipedia articles such as "See. Also" and "Links". This approach allows us to raise the performance of the algorithm and is designed for use in applications requiring high accuracy of the result is not, and better performance of the algorithm. These tasks include establishing a correspondence between the competencies and educational standard annotations disciplines of the curriculum or the task of analyzing the students' answers to the open questions in the form. The developed measure is cheap, reasonably accurate and accessible.

Keywords: *link, structure of article of Wikipedia, the database of Wikipedia, background knowledge, semantic similarity of concepts, concept, link graph, distance between concepts, count indexing, link-based Measure*

Введение

Одной из важных задач во многих процессах обработки естественного языка (автоматическое построение онтологий, информационный поиск, автоматическое реферирование и генерация аннотаций и др.) относится задача по определению семантической близости понятий. Действительно, в задачах автоматической обработки текстовой информации часто возникает необходимость определить, насколько сильно та или иная пара концептов (понятий) связана по смыслу, иначе говоря, оценить степень семантической близости между ними.

Под семантической близостью понимается смысловое расстояние между понятиями, которое задается на графе понятий-концептов. Понятия могут быть близки по смыслу, но не тождественны [2]. М.И. Варламов в работе [4] дает более точное определение: Мера семантической близости концептов – это числовая оценка степени их смысловой связанности.

К настоящему времени разработано достаточно много мер семантической близости-связанности понятий, в достаточно полной мере, представленные в [2-5,7]. В данной работе предлагается способ определения семантической близости понятий, основанный на применении метода WLM (Wikipedia Link-based Measure [7]) и меры близости по отдельным типам ссылок [3], с помощью которых вычисляется семантическая близость между терминами на основе ссылок, найденных в пределах соответствующих статей Википедии.

Фоновые знания

При определении семантической близости понятий необходимы дополнительные источники знаний. В одних случаях используются корпуса текстов, в других таксономия, тезаурус или онтология. Такие источники знаний считаются фоновыми знаниями, опреде-

DOI: 10.7256/2305-6061.2016.3.19560

ляющими принятое ограничение [7]. Корпуса текстов являются неструктурированными и неточными, а таксономии или тезаурусы, созданные в основном вручную, ограничены по своим масштабам.

Эти ограничения являются основной мотивацией в применении методов, основанных на использовании структуры и содержания Википедии.

Википедия, имея более одного миллионов статей и тысячи авторов, является быстро растущим, крупнейшим хранилищем знаний. Благодаря обширной сети перекрестных ссылок, порталов и категорий она также содержит большое количество явно определенной семантики. Это редкое сочетание масштаба и структуры делает Википедию привлекательным ресурсом для задач обработки естественного языка.

Таким образом, для определения семантической близости понятий в качестве фоновых знаний будет использовать базу данных Википедии. На второе июня 2016 года в русскоязычной Википедии насчитывалось 1 315 000 статей [5], в англоязычной – 5 168 199 статей [1].

Измерение связанности между понятиями

В отличие от других методов, основанных на Wikipedia, WLM [7] обеспечивает достаточно точные измерения, используя только ссылки между статьями, а не их текстовое содержание. В методе WLM используются внутритекстовые ссылки, и ссылки "См. также" (See also) и "Ссылки" (Links, External links). В данной работе проведем исследования, используя только ссылки из секций: "См. также" и "Ссылки". Для измерения связанности между понятиями построим ссылочный граф и выполним анализ его вершин.

Построение ссылочного графа понятия. Все страницы интернет динамически сгенерированные средствами JavaScript, JSP, PHP, ASP или разработанные другими веб-технологиями, основаны на HTML. Браузер разбирает HTML-код и отображает его в удобном виде. Наша задача получить и проанализировать этот код на предмет наличия в нем секций "См. также" и "Ссылки". Для выполнения парсинга HTML-кода нами использована библиотека JSoup.

Путь имеем два понятия x и y, необходимо определить меру их семантической близости-связности. Построим ссылочный граф как неориентированный граф $G = <V, E>$ с множеством вершин V и множеством ребер E.

Выполним запрос в Википедии на получение адреса статьи, описывающей понятие x. Осуществим парсинг статьи и находим секции, если они есть в наличии. Определяем множество V^1 ссылок первого уровня:

$$V^1 = V_1 \cap V_2,$$

где V_1 и V_2 есть множества ссылок из секций «См. также» и «Ссылки» соответственно.

Далее для каждого понятия первого уровня аналогично находим множество V^i, тогда множество ссылок i-го уровня определяется формулой:

DOI: 10.7256/2305-6061.2016.3.19560

$$V^i = \bigcap_{j=1}^{k} V_j^{i-1},$$

где k – количество ссылок на предыдущем уровне $(i - 1)$, $V_j^{\,i-1}$ – множество ссылок, соответствующих j-му понятию $(i - 1)$ -го уровня.

Аналогичным образом строится граф для второго понятия y. Для различения графов обозначим их, как $G_x =< V_x, E_x >$ и $G_y =< V_y, E_y >$.

Измерение связанности между понятиями. Определение семантической близости-связности понятий будем оценивать на основе расстояния (длины кратчайшего пути) между ними в ссылочном графе, основываясь на работе [3, с.1114]. При измерении могут возникнуть следующие ситуации:

1. ссылки на оба понятия x и y существуют в обоих ссылочных графах G_x и G_y ;
2. в ссылочных графах G_x и G_y отсутствует ссылка на второе понятие, но имеется ссылка на понятие z (одна или более), присутствующая в обоих графах;
3. в ссылочных графах G_x и G_y отсутствует ссылка на второе понятие и нет в наличии ссылок общих для обоих графов.

Если возникает третья ситуация, то будем считать, что понятия не связаны, поэтому исследовать будем первые две ситуации.

В работе [3] при вычислении расстояний между понятиями используется индексация графа ссылок Википедии. Как и в вышеназванной работе используем это понятие индекса графа с двухшаговым покрытием вершин (2-hop cover) при расчете связанности между понятиями. Рассмотрим расчеты при возможных ситуациях наличия понятий в ссылочных графах.

Первая ситуация. Назовем меткой $L(x)$ вершины $x \in V_x$ множество пар

$$L(x) = \{(y, dist_{G_x}(x, y))\}_{y \in C(x)}, C(x) \subset V_x,$$

где $dist_{G_x}(x, y)$ – расстояние между вершинами x, y в графе G_x. Индекс графа – такое множество меток его вершин, что какая-то пара вершин
$v_i, v_{i+1} \in V_x$ обе их метки содержат расстояние до вершины y на кратчайшем пути между ними.

Аналогично меткой $L(y)$ вершины $y \in V_y$ есть множество пар

$$L(y) = \left\{\left(x, dist_{G_y}(x, y)\right)\right\}_{x \in C(y)}, C(y) \subset V_y.$$

Если метки $L(x), L(y)$ обладают указанным свойством, расстояние между x и y может быть вычислено как

$$dist(x, y) = \min(dist_{G_x}(x, y), dist_{G_y}(y, x)). \qquad (1)$$

Вторая ситуация. Также как и в первом случае вычисляем метки $L(x)$ и $L(y)$. Отличие

DOI: 10.7256/2305-6061.2016.3.19560

заключается в том, что при вычислении меток будем определять расстояние от x до z и от y до z.

$$L(x) = \{(z, dist_{G_x}(x,z))\}_{z \in C(z)}, C(x) \subset V_x,$$

$$L(y) = \left\{\left(z, dist_{G_y}(y,z)\right)\right\}_{z \in C(y)}, C(y) \subset V_y.$$

$$dist(x,y) = dist_{G_x}(x,y) + dist_{G_y}(y,x). \quad (2)$$

Мера семантической близости между понятиями x,y можно определяется тем больше, чем меньше расстояние между ними, тем выше мера:

$$sim(x,y) = \frac{1}{dist(x,y)}. \quad (3)$$

Пример определения семантической близости между понятиями.

Пусть x = "Вейвлет", y = "Цифровая обработка сигналов".

Для построения графов G_x, G_y, представляющих собой дерево пространства состояний, и для реализации поиска совпадающих ссылок x_i, y_i в графах G_x, G_y в работе использован один из слепых методов перебора метод поиска в ширину.

Построенные графы представлены в виде матриц инцидентности, примеры матриц приведены на рис. 1, 2. Граф G_x построен для понятия «Вейвлет», а граф G_y – «Цифровая обработка сигналов».

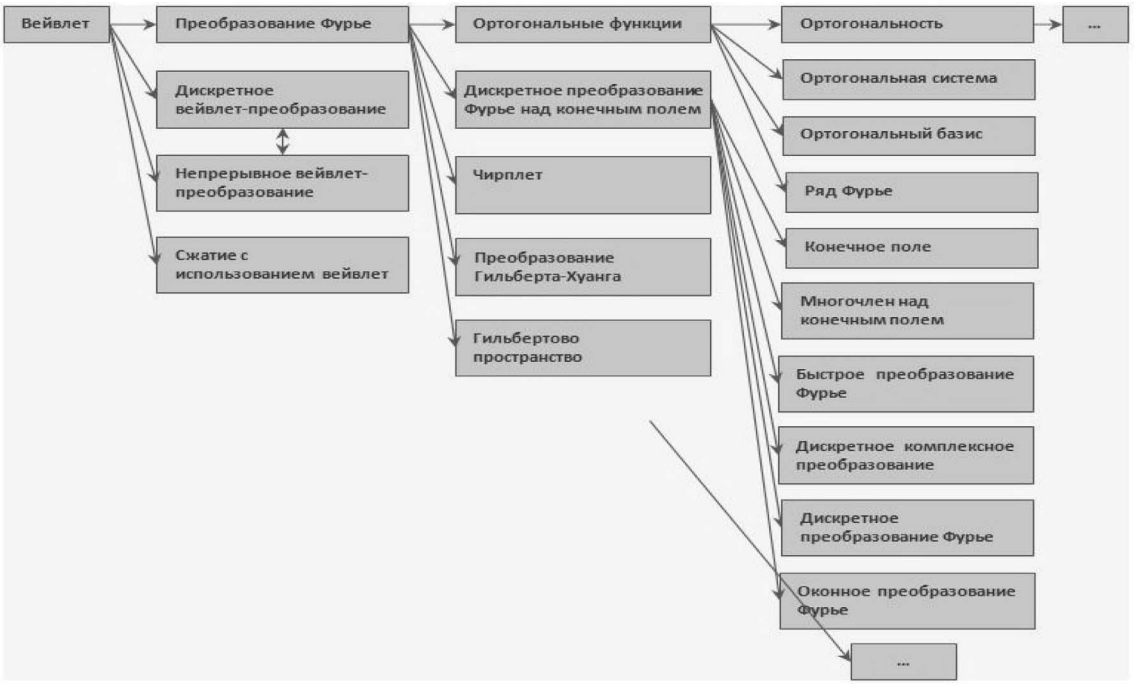

Рисунок 1 – Пример фрагмента ссылочного графа понятия «Вейвлет»

DOI: 10.7256/2305-6061.2016.3.19560

Как показывают рисунки, данным понятиям соответствует вторая ситуация. Рассчитаем меру семантической близости этих понятий.

$$L(x) = \{(\text{"Преобразование Фурье"}, 1)\},$$
$$L(y) = \{(\text{"Преобразование Фурье"}, 1)\}_{z \in C(y)}.$$

Так как понятия x и y находятся во второй ситуации, т.е. они связаны посредство третьего понятия "Преобразование Фурье", то расстояние между ними вычисляется по формуле (2):

$$dist(x, y) = dist_{G_x}(x, y) + dist_{G_y}(y, x) = 1 + 1 = 2.$$

Мера семантической близости между понятиями x,y определяется тем больше, чем меньше расстояние между ними:

$$sim(x, y) = \frac{1}{dist(x, y)} = \frac{1}{3} = 0{,}5.$$

Проведем дополнительные эксперименты. Найдем семантическую близость между несколькими понятиями предметной области «Искусственный интеллект»: алгоритм Rete, база знаний, дерево принятия решений, машина вывода, экспертная система. Результаты вычислений показаны в табл. 1.

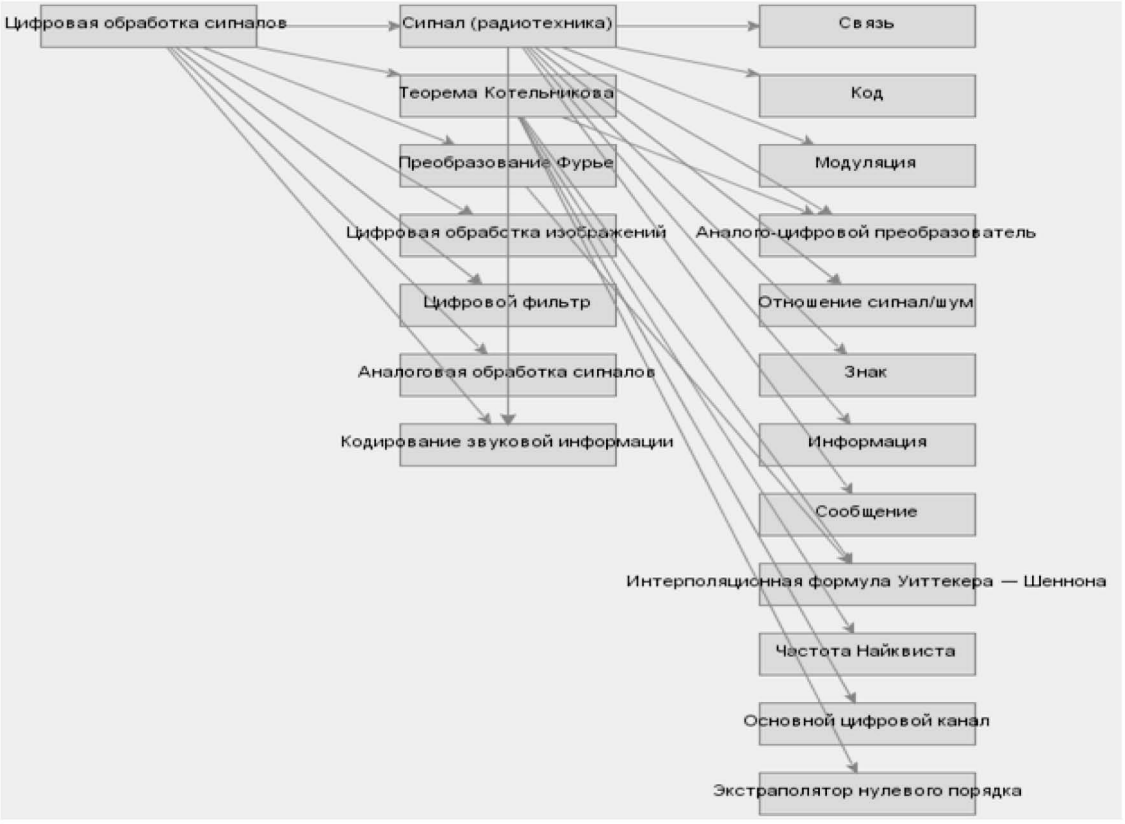

Рисунок 2 – Пример ссылочного графа понятия «Цифровая обработка сигналов»

DOI: 10.7256/2305-6061.2016.3.19560

Таблица 1 – Меры семантической близости понятий

N	Понятие		Мера семантической близости
	x	y	$sim(x, y)$
1	2	3	4
1	Алгоритм Rete	База знаний	0,25
2		Машина вывода	0,50
3		Дерево принятия решений	0
4		Экспертная система	0,50
5	Дерево принятия решений	База знаний	0
6		Алгоритм Rete	0
7		Экспертная система	0,25
8		Машина вывода	0
9	Экспертная система	База знаний	1,00
10		Алгоритм Rete	0,50
11		Машина вывода	0,5
12		Дерево принятия решений	0,25

Далее были определены диапазоны значений мер сильносвязанных, среднесвязанных и слабосвязанных понятий. Для этого использовался метод экспертной оценки. В результате установлены следующие диапазоны значений: мера близости сильносвязанных понятий изменяется в диапазоне от 0.5 до 1; среднесвязанных – от 0.2 до 0.49 и мера семантической близости слабосвязанных понятий имеет значения меньшие 0.2.

Заключение

В работе предложен и исследован способ определения меры семантической близости между понятиями. Основное отличие данного способа от других заключается в использовании простых ссылок структуры гиперссылок Википедии. Этот подход предлагает меру более дешевую, точную и доступную меру, чем ранее разработанные меры. Дешевизна меры объясняется тем, что обширное текстовое содержание Википедии в значительной степени может быть проигнорировано. Более высокая точность меры связана с тем, что она использует семантику понятий которая определена вручную. Доступность объясняется применением в качестве фоновых знаний Википедии.

Библиография :

1. Английская Википедия [Электронный ресурс]. – URL: https://ru.wikipedia.org/wiki/Английская_Википедия (дата обращения: 20.06.2016).

DOI: 10.7256/2305-6061.2016.3.19560

2. Анисимов А.В. Метод вычисления семантической близости-связности между словами естественного языка / А.В. Анисимов, А.А. Марченко, В.К. Кисенко // Кибернетика и системный анализ. 2011. № 4. С.18-27.
3. Варламов М.И. Расчет семантической близости концептов на основе кратчайших путей в графе ссылок Википедии / М.И. Варламов, А.В. Коршунов // Машинное обучение и анализ данных. 2014. Т. 1. № 8. С. 1107-1125.
4. Варламов М.И. Расчет семантической близости концептов на основе кратчайших путей в графе ссылок Википедии [Электронный ресурс]: презентация / М.И. Варламов, А.В. Коршунов // URL: www.machinelearning.ru/wiki/images/f/fd/Varlamov2014iip.pdf (дата обращения: 20.06.2016).
5. Русская Википедия [Электронный ресурс]. – URL: https://ru.wikipedia.org/wiki/Русская_Википедия (дата обращения: 20.06.2016).
6. Турдаков Д.Ю. Texterra: инфраструктура для анализа текстов / Д.Ю. Трудаков и др. // Труды Института системного программирования РАН. 2014. Т. 26. Вып. 1. С. 421-438.
7. Witten I., Milne D. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links // Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA. 2008. P. 25-30.

References:

1. Angliiskaya Vikipediya [Elektronnyi resurs]. – URL: https://ru.wikipedia.org/wiki/Angliiskaya_Vikipediya (data obrashcheniya: 20.06.2016).
2. Anisimov A.V. Metod vychisleniya semanticheskoi blizosti-svyaznosti mezhdu slovami estestvennogo yazyka / A.V. Anisimov, A.A. Marchenko, V.K. Kisenko // Kibernetika i sistemnyi analiz. 2011. № 4. S.18-27.
3. Varlamov M.I. Raschet semanticheskoi blizosti kontseptov na osnove kratchaishikh putei v grafe ssylok Vikipedii / M.I. Varlamov, A.V. Korshunov // Mashinnoe obuchenie i analiz dannykh. 2014. T. 1. № 8. S. 1107-1125.
4. Varlamov M.I. Raschet semanticheskoi blizosti kontseptov na osnove kratchaishikh putei v grafe ssylok Vikipedii [Elektronnyi resurs]: prezentatsiya / M.I. Varlamov, A.V. Korshunov // URL: www.machinelearning.ru/wiki/images/f/fd/Varlamov2014iip.pdf (data obrashcheniya: 20.06.2016).
5. Russkaya Vikipediya [Elektronnyi resurs]. – URL: https://ru.wikipedia.org/wiki/Russkaya_Vikipediya (data obrashcheniya: 20.06.2016).
6. Turdakov D.Yu. Texterra: infrastruktura dlya analiza tekstov / D.Yu. Trudakov i dr. // Trudy Instituta sistemnogo programmirovaniya RAN. 2014. T. 26. Vyp. 1. S. 421-438.
7. Witten I., Milne D. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links // Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA. 2008. R. 25-30.

DOI: 10.7256/2305-6061.2016.3.19560