OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

АЛГОРИТМ ВЫДЕЛЕНИЯ БИНАРНЫХ ОТНОШЕНИЙ В ТЕРМИНОЛОГИЧЕСКОЙ СЕТИ

Тузикова А.В.

Факультет вычислительной математики и кибернетики,
Московский государственный университет имени М.В. Ломоносова,
г. Москва, Россия

kabaylova@yandex.ru

В работе рассматривается возможность частичной автоматизации технологии построения терминологических сетей. Предлагается метод выявления семантических связей, основанный на использовании метрики редактирования для вычисления расстояний между понятийными вершинами. Работа алгоритма иллюстрируется на модельной терминологической сети.
Ключевые слова: терминологическая сеть; понятие; метрика редактирования.

Введение

В современном мире ощущается потребность в структурном представлении данных, удобном для их исследования. Одним из способов структуризации данных является "Универсальное терминологическое пространство" (УТП) [Мальковский и др., 2002], проект которого подразумевает построение семантической сети из проблемно-ориентированных взаимосвязанных глоссариев. Развитие УТП обеспечивается редакторами-экспертами, пополняющими терминологическую сеть путем обработки глоссариев, выделения понятий и установления бинарных семантических связей.

Проблема качественного редактирования терминологической сети порождает задачу автоматизации этой деятельности.

1. Введение в терминологические сети

Терминологическая сеть [Мальковский и др., 2012] относится к классу семантических сетей [Sofa, 1992] и представляет собой граф, вершинами которого являются определения терминов, а ребрами – бинарные отношения между ними. Фактически, вершина представляет собой структуру данных, содержащую информацию о термине: синонимы, определение, а так же дополнительные сведения о термине или о связанных с ним вершинах. Ребрам графа соответствуют два допустимых типа отношений: “это-есть” и “относится-к”, к которым сводятся все остальные бинарные отношения.

При добавлении новых данных в существующую терминологическую сеть редактор обрабатывает определенное количество новой информации, постоянно взаимодействуя с сетью значительных размеров. Для облегчения деятельности редактора предлагается алгоритм выделения потенциально возможных связей между включаемой в сеть понятийной вершиной и вершинами расширяемой терминологической сети.

2. Алгоритм автоматического выявления связей

В основу алгоритма выявления связей положено предположение о наличии в определении или в словарной статье достаточной информации для определения возможных связей с понятийными вершинами терминологической сети.

Исходное определение можно рассматривать как строку (см. раздел 2.1). Понятийную вершину так же можно рассматривать как особым образом структурированную строку. Рассматривая определение и понятийную вершину как строки, можно определить расстояние между ними с использованием метрики редактирования [Деза и др., 2008]. В свою очередь, если рассматривать строку как набор слов, тогда в общем наборе операций редактирования достаточно рассматривать лишь операции замены и вставки-удаления символов, что соответствует метрике Левенштейна [Левенштейн, 1965].

Метрика Левенштейна применяется неоднократно для вычисления расстояний между парами слов из сравниваемых строк. Для принятия решения о значимости вычисленного расстояния используется порог, который задается величиной *Val*, устанавливаемой опытным путем и являющейся единственным параметром при определении близости понятийных вершин.

Таким образом, задача выявления связей сводится к задаче определения расстояний между строками. Упомянутое расстояние предлагается вычислять как:

$$Dist = \sum_{succNum} 2^{k_{succNum}} + \frac{imp}{10^2} - \frac{unimp}{10^5}. \qquad (1)$$

$$k_{succNum} = Val - \max\{1, pDist\}. \qquad (2)$$

В формуле (1) суммирование ведется по всем новым расстояниям, являющимся существенными в соответствии с порогом Val; imp – число пар слов, расстояние между которыми влияет на результат сравнения; unimp – число пар слов, расстояние между которыми не влияет на результат сравнения; 102 и 105 – постоянные коэффициенты, выбранные таким образом, чтобы можно было свети в одно число количество существенных результатов, количество несущественных результатов, а также вычисленное с использованием метрики Левенштейна расстояние; pDist – вычисленное значение расстояния для пары слов. Взятие максимума позволяет рассматривать идентичные слова как просто близкие, что дает возможность комбинациям слов конкурировать с совпадающими словами при вычислении их вклада в расстояние Dist. Следует отметить, что ksuccNum ≥ 0, так как к моменту вычисления гарантируется существенность значения переменной pDist.

Предложенный подход к определению близости двух понятийных вершин позволяет учесть не только влияние результатов отдельных сравнений, но и их качество. Так, словосочетания “свободные места” и “свободными местами” окажутся более близкими, чем “свободные места” и “свободные стулья”, что соответствует сути дела. Если же установить значение параметра *Val* = 0 (что соответствует абсолютной идентичности сравниваемых слов), сложится обратная ситуация.

В силу выбора упомянутого числового коэффициента 10^5, число несущественных сравнений оказывает влияние лишь на порядок перечисления выявленных кандидатов, предлагаемых редактору.

2.1. Входные и выходные данные

Алгоритм выявления связей позволяет модифицировать первоначально заданную терминологическую сеть. В качестве входных данных алгоритм использует выделенное пользователем определение. Приведем типичные примеры исходных определений, взятых из музыкального словаря [Риман, 2008] и в модифицируемой сети не представленных.

Пример 1

Бассанелло – ныне вышедший из употребления деревянный духовой инструмент, родственный фаготу. У него был двойной язычок, заключенный в воронкообразный мундштук и изогнутая шейка (в виде буквы S). Бассанелло делался трех различных величин (басовый, теноровый и дискантовый). В старинных органах были также регистры Bassanelli.

Пример 2

Чибыза (чебызга) – дудка из камыша или из дерева; народный музыкальный инструмент киргиз-кайсаков.

Пример 3

Магадис – струнный инструмент древних греков, похожий на арфу, число струн коего доходило до 40; одно место в псевдо-аристотелевых проблемах указывает на то, что на магадисе играли мелодию октавами.

Выходными данными алгоритма являются термины, соответствующие понятийным вершинам терминологической сети, которые могут быть связаны с новой понятийной вершиной, формируемой на основе введенного определения, связью одного из типов "это-есть" и "относится-к".

2.2. Идея алгоритма выявления связей

Алгоритм выявления связей состоит из двух этапов.

Этап 1. [Определение расстояний]. Определить расстояния между введенным термином, его определением и соответствующей информацией из понятийных вершин с помощью метрики Левенштейна. Сравнения элементов, представленных в виде строк, следует провести между парами “определение” – термин, соответствующий понятийной вершине; и “термин” – информация, полученная из понятийной вершины. Остальными способами комбинирования пренебречь вследствие малой вероятности возникновения связей "это-есть" и "относится-к" между практически идентичными понятиями.

Этап 2. [Добавление кандидатов]. Добавить подходящие вершины, а так же вычисленные для них значения к числу кандидатов на добавление связей. Новые “варианты” во множество кандидатов добавлять лишь взамен самого худшего из текущих.

По результатам выполнения этапов 1 и 2 формируется набор из ограниченного постановкой конкретной задачи числа кандидатов. После обработки всех вершин сети работа алгоритма завершается, результатом является набор понятийных вершин терминологической сети,

которые могут быть связаны отношениями “это-есть” или “относится-к” с новой, сформированной на основе предложенных для обработки данных, понятийной вершиной.

После завершения поиска кандидатов полученные результаты предоставляются редактору для принятия и документирования в сети соответствующих решений, так же как информация о предлагаемых понятийных вершинах. Таким образом, алгоритм облегчает выявление связей между вершинами, хотя и оставляет задачу выбора редактору.

Результатом взаимодействия пользователя с системой является расширенная терминологическая сеть, в состав которой включены сформированная вершина и выбранные для добавления связи.

Следует отметить, что алгоритм не предусматривает усечение или нормализацию слов, то есть сравнения могут проводиться между различными формами одного и того же слова.

3. Применение к конкретной предметной области

Для иллюстрации работы метода выявления связей используется сеть, соответствующая области “Музыка”, содержащая 847 понятийных вершин и входящая в состав glossary.ru [Мальковский и др., 2002].

Для модельной области проведено расширение терминологической сети включением в нее новых понятийных вершин. В таблицах 1-3 приведены полученные численные результаты для небольшого числа новых данных, добавленных в сеть с использованием алгоритма. Входные данные приведены в примерах 1-3. В таблице 4 предложены к рассмотрению результаты анализа данных из таблиц 1-3 редактором-экспертом.

В таблицах 1-3 столбец "*Dist*" содержит вычисленное значение семантической близости введенного определения и соответствующей понятийной вершины; в столбце “Вершины” приведены по два предлагаемых варианта для рассматриваемых входных данных.

Таблица 1 – Расчетные значения расстояний для модельной области “Музыка” для термина “бассанелло” – *Пример 1*

Вершина	*imp*	*unimp*	*Dist,* *до* 10^{-3}
Двойной язычок	2	80	16.019
Духовые музыкальные инструменты	2	121	12.019

Для термина “бассанелло” выбор вершины “двойной язычок” обусловлен наличием соответствующего словосочетания в тексте определения в именительном падеже, что делает его более ценным с точки зрения алгоритма, чем понятие “духовые музыкальные инструменты”.

Таблица 2 – Расчетные значения расстояний для модельной области “Музыка” для термина “чибыза” – *Пример 2*

Вершина	*imp*	*unimp*	*Dist,* до 10^{-3}
Народные музыкальные инструменты	3	30	24.030
Духовые музыкальные инструменты	2	31	16.020

Для термина “чибыза” можно наблюдать аналогичную ситуацию. Выбор вершины “Народные музыкальные инструменты” основан на непосредственном наличии в тексте определения словосочетания “народный музыкальный инструмент”. “Духовые музыкальные инструменты” имеют меньшее значение в связи с меньшим числом слов в названии вершины, встречающихся в тексте определения.

Таблица 3 – Расчетные значения расстояний для модельной области “Музыка” для термина “магадис” – *Пример 3*

Вершина	*imp*	*unimp*	*Dist,* до 10^{-3}
Струнные музыкальные инструменты	2	79	16.019
Струнные щипковые музы-кальные инструменты	2	106	16.019

Для термина “магадис” имеет место влияние количества информации (числа слов) в понятийных вершинах на упорядоченность результатов в соответствии с (1).

В таблице 4 обозначения “О” и “Э” соответствуют типам связей “относится-к” и “это-есть”.

Таблица 4 – Результаты работы эксперта-редактора

Обрабатывае-мые данные	Вершина	Тип связи	Выбор
Пример 1	Духовые музыкальные инструменты	Э	Система
Пример 2	Народные музыкальные инструменты	Э	Система
	Духовые музыкальные инструменты	Э	Система
Пример 3	Древнегрече-ская	О	Редактор

	культура		
	Струнные музыкальные инструменты	Э	Система

При анализе результатов, предлагаемых алгоритмом, для выборки в 40 определений новых терминов, 72,5% случаев были оценены как удовлетворительные и не требующие глубокого вмешательства редактора-эксперта. Таким образом, работа редактора при использовании представленного алгоритма в большинстве ситуаций сводится к выбору одного из предложенных вариантов.

Заключение

Предложен метод предварительной настройки терминологической сети, основанный на использовании метрики редактирования для выделения потенциальных связей между новыми данными и вершинами сети.

Использование метрики редактирования позволяет свести процесс поиска связей между понятийными вершинами к определению расстояния между строками.

Описанный метод дает релевантные результаты и позволяет за счет частичной автоматизации снизить нагрузку на редакторов терминологической сети при ее расширении и повысить объективность.

Предложенный подход может быть использован в качестве основы при создании программы для введения в сеть новых данных порциями большего объема.

Библиографический список

[Деза и др., 2008] Деза Е.И., Деза М.-М. Энциклопедический словарь расстояний / Елена Деза, Мишель-Мари Деза – пер. с англ. В.И. Сычева; Моск. гос. пед. ун-т ; Нормальная высш. шк., Париж. – М.: Наука, 2008. – с. 178-186.

[Левенштейн, 1965] Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии Наук СССР, т. 163, №4 – М.: Наука, 1965. – с. 845–848.

[Мальковский и др., 2002] Мальковский М.Г., Соловьев С.Ю. Универсальное терминологическое пространство. // Труды Международного семинара Диалог'2002 "Компьютерная лингвистика и интеллектуальные технологии", т.1. М.: Наука, 2002, с. 266-270.

[Мальковский и др., 2012] Мальковский М.Г. Терминологические сети / М.Г.Мальковский, С.Ю.Соловьев // OSTIS-2012. Материалы конференции. С. 77-82**.**

[Риман, 2008] Риман Г. Музыкальный словарь [Электронная версия] / Пер.: Б. Юргенсон. - М.: Директмедиа Паблишинг, 2008. - 10440 с.

[Sofa, 1992] John F. Sowa Semantic Networks / Encyclopedia of Artificial Intelligence, second edition, Wiley, New York, 1992.

ALGORITHM OF BINARY RELATIONS EXTRACTION FROM TERMINOLOGICAL NETWORK

Tuzikova A.V.

Lomonosov MSU CS department, Moscow, Russia

kabaylova@yandex.ru

In the article the possibility of partial automatization of terminological networks forming process is considered. The algorithm for identification semantic relations is proposed. It is based on using editing metrics for determining distances between concept nodes of the terminological network.

Introduction

There is lot of unstructured information in the world around us nowadays. It is necessary to make it structured. One of the ways to do that is the universal terminological space (UTS). This project implies creation of the semantic network over the set of problem-oriented glossaries. In its semantic network two types of relations are used, they are “it-is” and “belongs-to”. The extension of the existing network by including new concept nodes and binary relations in it makes it necessary to automate the activity of scientific editors. As a result, the solution described at the article appeared.

Main Part

The main idea of the suggested algorithm is to find distances between the concept node, created by processing new data, and nodes of the existing terminological network that is being extended. The output of the algorithm is the information about several nodes that can be connected with the new one.

The algorithm is based on using editing metrics. The Levenshtein distance was chosen.

Let’s think about new data and existing nodes as about strings. Then let’s look at strings (that obviously are sentences) as a set of words. That approach allows to calculate distances between new data and nodes as distances between separate words that are summed by several rules to find required value.

In the article there are results concerning the data connected with the Music area.

Conclusion

The proposed method allows to find semantic relations between new data and nodes of the extensible terminological network. Thus, it contributes to decrease loading on the editors and to increase the objectivity.