УДК 004.89

CONCEPT MAP MINING: МЕТОД СОЗДАНИЯ КОНЦЕПТУАЛЬНЫХ КАРТ

CONCEPT MAP MINING: CREATING CONCEPT MAPS METHOD

Г.А. Евсеев, Н.Д. Тодосиев
Москва, МГТУ им. Н.Э. Баумана
G.A. Evseev, N.D. Todosiev
Moscow, BMSTU

Аннотация. *В статье описывается метод Concept map mining для автоматизированного создания концептуальных карт. Приведены определения Концептуальной карты, рассмотрены области, в которых используется концептуальная карта, как способ представления знаний. Представлен пример концептуальной карты. Описан принцип построения концептуальных крат. Так же кратко рассмотрены разные подходы к реализации метода Concept map mining. Поэтапно рассмотрена и описана реализация Concept map mining на основе неструктурированного текста на Русском языке. Так же поднимается проблема оценки созданной концептуальной карты, сложность создания метрики, субъективность оценивания концептуальных карт, созданных с помощью полностью автоматизированных методов. А также предлагается метрика, решающая данную проблему, основанная на принципе "Золотого стандарта".*

Ключевые слова*: Концептуальная карта, автоматическая генерация концептуальной карты, анализ текста.*

Abstract. *This article describes the concept map mining method for automated creation of concept maps. The definition of the Concept map is given, the areas in which the concept map is used is considered as a way of representing knowledge. An example of a conceptual map construction is described. Different approaches to the implementation of this method are also briefly considered. The Concept map mining implementation based on unstructured text in the Russian language is reviewed and described in stages. he problem of evaluating the created conceptual map, the complexity of creating a metric, the subjectivity of evaluating conceptual maps created using fully automated methods also rises. And also a metric is proposed that solves this problem, based on the principle of the "Gold Standard".*

Keywords*: Concept map, automatic concept map generation, text analysis.*

Введение

В настоящее время наблюдается рост использования концептуальных карт. Самое частое применение концептуальных карт - сбор и представление уже полученных знаний в форме, удобной для восприятия. Более того, концептуальные карты известны как эффективный инструмент для организации и навигации по большим объемам информации.

Так же концептуальные карты могут быть использованы как средство для представления плана обучения, где концепты будут представлять набор целей, а связи между ними – средствами для достижения этих целей. Зачастую сложно с нуля составить карту по выбранной теме. Каркас схемы, предоставленный экспертом в интересующей области, может сильно облегчить задачу, но при

индивидуальном изучении зачастую тяжело найти такого человека. Поэтому информационная система, которая заменит эксперта и предоставит каркас концептуальной карты может быть очень полезна в такой ситуации.

Концептуальная карта – это графический инструмент, использующийся для структурирования и представления знаний. Она включает в себя концепты, которые чаще всего представлены существительными и именными группами и связями между ними, указываемые линией, соединяющей два концепта. Обозначение линии глаголом или глагольной группой создает цепочку концепт-связь-концепт, которую можно прочитать как предложение. Эта цепочка называется утверждением [1]. Пример концептуальной карты представлен на рисунке.

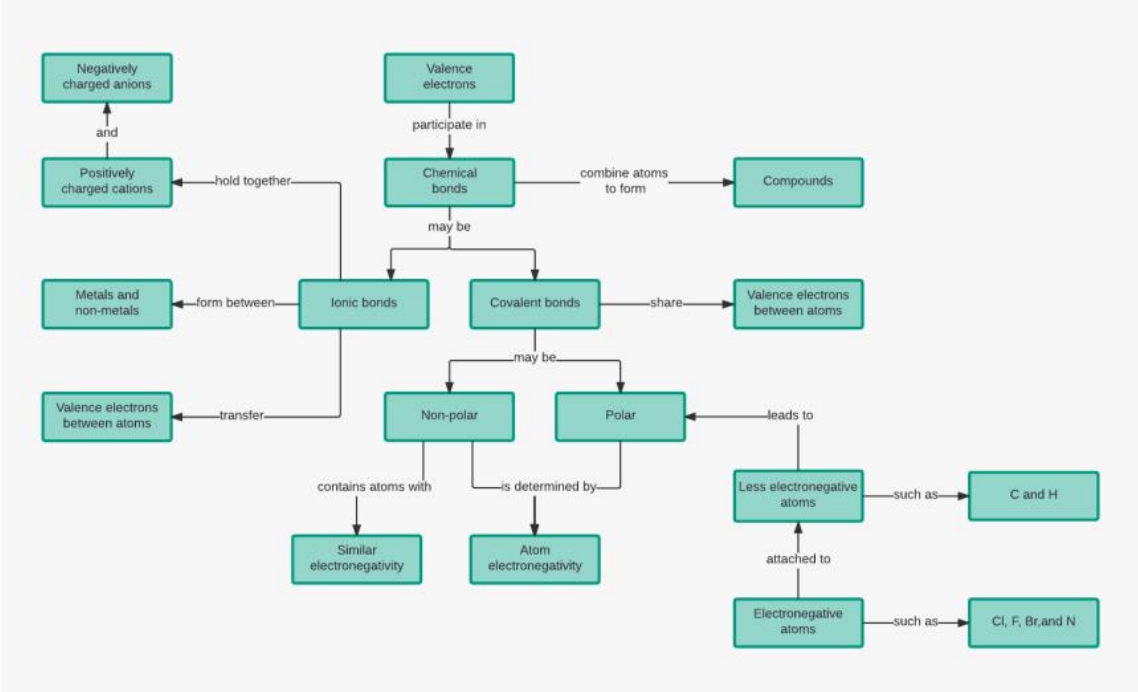

Рис. 1. Пример представления знаний в виде концептуальной карты.

Автоматическая генерация концептуальных карт из документов называется Concept Map Mining (CMM) [2]. CMM может быть как полуавтоматическим, так и полностью автоматизированным. В полуавтоматическом режиме система находит и предлагает некоторые элементы карты, и человек должен закончить карту вручную, используя предоставленную информацию. В автоматической реализации, на выходе пользователь получает конечную концептуальную карту исходя из входных данных.

В данной статье представлен метод CMM в общем виде и описан метод создания концептуальной карты из неструктурированного текста на Русском языке.

Concept Map Mining

Семантическая сеть представляет собой структуру для представления знаний в виде шаблона взаимосвязанных вершин и дуг [3]. Концептуальная карта является особым типом семантической сети, гибкой и ориентированной на человека. Она представлена в виде направленного графа, где узлы представляют концепты, а дуги - связи между ними [4].

Понятие концептуальной карты приписывается к учению теоретика Джозефа Новака, где его группа исследователей описала процесс обучения человека как пожизненный процесс усвоения новых концептов и связей между ними в личную концептуальную структуру [1]. Новак адоптировал модель семантической сети и создал понятие концептуальной карты, как инструмент для графического представления концептуального понимания информации обучающимся в определенной области. Его изначальной идеей было, что концептуальная карты должна быть нарисована рукой обучающимся после определения всех главных идей и их классификации в иерархической манере. Топология концептуальных карт может принимать разные формы от иерархической до неиерархической и формами, управляемые данными.

В общем виде, иерархическая концептуальная карта может быть определена как набор из концептов (C), связей (R) и уровень обобщения (L). Каждый концепт (c_i) – это слово или фраза, уникальная для набора концептов. Каждая связь (r_i) соединяет два концепта и обладает меткой связи, которая определяет тип связи между этими двумя концептами. Уровень обобщения (l_i), который является набором концептов, которые разделяют один и тот же уровень обобщения в концептуальной карте.

Concept map mining представляет собой процесс извлечения информации из одного или более документов для автоматического создания концептуальной карты. Созданная карта является обобщённой сводкой исходного текста [2].

Со точки зрения CMM, документ может быть представлен как набор концептов (C_d) и связей (R_d). Тремя основными фазами Concept map mining являются: извлечение концептов, извлечение связей, обобщение. Первый этап является определением и извлечением всех концептов, которые представлены субъектами и объектами в тексте – обычно это существительные и именные фразы [5]. Когда известна синтаксическая и семантическая связь между концептами, то возможно извлечь связь между ними, что является целью второго этапа. Последним этапом является обобщение документа, где на выходе мы получим набор иерархически выстроенных триплетов концепт-связь-концепт. Концептуальная карта предназначена для анализа человеком, поэтому желательно чтобы она не имела слишком много концептов. Так же важно, чтобы в концептуальной карте использовались термины, которые использовались в изначальном документе.

Первые наброски метода CMM были представлены в работе Трочима, который использовал статистический анализ для решения данной задачи. Группа людей во время собрания накидывала некоторое количество утверждений, которое относилось к теме собрания. Каждый участник оценивал все утверждения, создавая индивидуальную матрицу. После все матрицы суммировались в групповой аппроксимирующий массив. Наиболее важные утверждения выбирались с метода многомерного шкалирования. Такой подход и сейчас используется в CMM [6].

Подходы использующиеся в concept map mining

Процесс Concept map mining может быть выполнен методами обработки естественного языка, такие как извлечение информации (IE – Information

Extraction), информационный поиск (IR – Information Retrieval) и автоматическое суммирование (Automatic summarization). IE – это задача автоматического извлечения структурированных данных, таких как сущности и связи из неструктурированных или слабоструктурированных машиночитаемых документов. IR представляет собой процесс поиска неструктурированной информации, удовлетворяющий информационным потребностям, а автоматическое суммирование является процессом сокращения набора данных для создания подмножества, содержащего наиболее важную информацию в исходном содержании [7]. Результатом автоматического суммирования могут быть выдержка или реферат. Выдержка является подмножество текста, содержащего наиболее важную информацию, выделенную из оригинала без изменений, в то время как рефератом является перефразированная выдержка из текста [8].

Методы, использующиеся в этих областях, являются статистическими методами на основе правил и методами машинного обучения. Относительно недавно появился интерес к объединению конечных автоматов с моделями условной вероятности, как например модели энтропии Маркова и условно случайные поля [7]. Большинство классических методов суммирования также являются числовыми и основаны на модели взвешивания, такими, как например term frequency-inverse document frequency (TF-IDF). Методы машинного обучения часто обеспечивают точное извлечение на основе классификации, использующей бинарную или нечеткую логику. Такие методы могут быть использованы как основной метод или в гибридных системах для обеспечения ресурсами других процессов. Современные подходы включают гибридные подходы с использованием алгоритмов в комбинации со сторонними наборами данных [8], [9], суммированием, основанным на нечеткой логике и роевом интеллекте [10]. В области обработки естественного языка, численные методы могут быть обогащены словарями терминов [7]. Но существует проблема со словарями: они должны быть заранее созданы для определенной области. Так же ограничивающий фактор использования лингвистических методов – это отсутствие нужных методов и инструментов для многих языков.

Реализация CMM на основе неструктурированного текста на Русском языке

CMM создает концептуальные карты из неструктурированных текстов используя статистические и data mining техники, обогащенные лингвистическими средствами. Одна концептуальная карта создается из одного документа. Главные шаги процедуры показаны на рисунке.

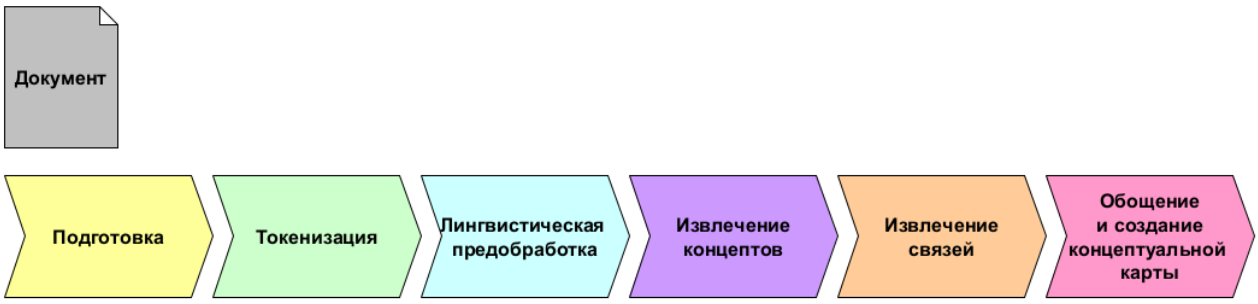

Рис. 2. Этапы процедуры CMM из текстов на Русском языке.

Во время этапа подготовки текст извлекается из документа. Все элементы, не несущие в себе никакой смысловой нагрузки, например сноски на список литературы, разметка удаляются и сохраняется очищенный текст.

На этапе токенизации, обработанный документ разделяется на предложения, каждое предложение токенизируется – выделяются базовые элементы языка. Этими элементами чаще всего являются слова или фразы разделенные не буквенно-цифровыми токенами, такие как пробел или пунктуационный знак. Все слова будут преобразованные в нижний регистр. Будут исключения, особенно для аббревиатур, знаков переноса, цифр, и некоторых других сущностей. Многие имена собственные образованы от имен нарицательных, и отличаются только регистром. Будет использован простой алгоритм, который будет приводить слова в начале предложения к нижнему регистру, а слова начинающиеся с заглавной буквой в середине предложения оставлять неизменными. Это потребуется для того, чтобы при отборе концептов имя нарицательное и идентичное ему имя собственное не определялись как одна сущность. На основе текста будет создана таблица связи аббревиатур и их расшифровки, и в тексте будут заменены все аббревиатуры на их расшифрованную фразу. При токенизации важно учитывать связь слов и предложений, потому что последующий анализ будет использовать эту информацию. Для токенизации будет использован конвейер, взятый из библиотеки spacy, обученный на модели ru_core_news_lg [11].

Следующий этап – это лингвистическая предобработка. В начале этой фазы из набора токенов будет убраны все шумовые слова, и все слова будут нормализованы. Нормализация слов может быть произведена с помощью стемминга или лемматизации [12]. Полученные слова будут размечены по частям речи. Для лемматизации и частеречной разметки будет так же использован конвейер, взятый из библиотеки spacy, обученный на модели ru_core_news_lg.

Этап номер четыре – извлечение концептов. На данном этапе будет создана двоичная матрица кандидатов в концепты, на основе частеречной разметки. В ней будут находится все концепты, которые встречаются в тексте, номер строки является номером предложения, а номер столбца отдельным концептом. В общем случае, субъект в предложении представляется первый концепт, а вторым концептом является объект этого же предложения. Такая матрица так же упростит подсчет такого индекса, как TF-IDF index, который определяет частоту, с которой каждый концепт встречается в тексте. Вычисление этого индекса

понадобится на последнем этапе, для определения вершины концептуальной карты.

На этапе извлечения связей, между всеми концептами будут определены семантические связи. Тип и метка связи могут между двумя концептами в простом предложении определяется основным глаголом. Для этого этапа используется модель re_rured. Модель основана на модели Адаптивного порогового значения и локализованного группирования контекста с добавлением NER сущностей в качестве дополнительного входного данного [13]. Так же потребуется таблица связи меток, которая будет создана ранних этапах, чтобы связать метки, которые будут выданы при частеречной разметке с метками, которые принимает на вход данная модель. На выходе мы получим код связи между двумя концептами в каждом предложении и его тип. Далее на основе двух концептов и их связи будут создана цепочка концепт-связь-концепт, которая называется предложение. Для каждого набора концептов будет создано предложение.

Концептуальная карта, которая дает представление о содержании документа, с минимальной избыточностью является результатом этапа обобщения и создания концептуальной карты. На данном этапе на основе TF-IDF индекса будет посчитан коэффициент важности предложения и на основе этого коэффициента будет определится расположение предложений в иерархии концептуальной карты. Предложение с наивысшим показателем коэффициента важности предложения будет помечено как начальное.

Оценка точности созданной концептуальной карты

Оценка автоматически созданной концептуальной карты является очень сложной задачей, которая работает со знаниями и значит, что она очень субъективна. Оценка будет производится с использованием концептуальных карт экспертов, которые будут приняты как “золотой стандарт”, который используется для автоматической оценки эссе – тоже очень субъективной и комплексной задачи. Случайно выбирается набор документов, по которым от двух до четырех экспертов создадут концептуальные карты. Далее будет посчитана human inter-annotator agreement метрика, так же будет посчитана machine-human agreement метрика. После подсчета проведется сравнение двух показателей, и если показатель human-machine agreement больше или равен human inter-annotator agreement, то производительность системы допустима [14].

Заключение

В статье были кратко описаны области применения концептуальных карт и объяснена важность автоматизации их создания. Был рассмотрен метод для автоматизированного создания концептуальных карт: concept map mining, а также кратко рассмотрены разные подходы к реализации данного метода. Была описана реализация прототипа CMM на основе неструктурированного текста на Русском языке и предложено решение проблемы об оценки точности полученной концептуально карты.

Список литературы

1. Novak J.D., Cañas A.J., The theory underlying concept maps and how to construct and use them //IHMC CmapTools 2006-01, Rev 01-2008, Florida Institute for Human and Machine Cognition, Pensacola Fl, Tech. Rep., January 2008.
2. Villalon J.J., Calvo R.A., Concept map mining: A definition and a framework for its evaluation //Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on, vol. 3, pp. 357–360, 2008.
3. Kramer B.M., Mylopoulos J. Representation, knowledge, Encyclopedia of Artificial Intelligence //Wiley-Interscience Publication, 1987, vol. 2, pp. 206–214.
4. McNeese M.D., Zaff B.S., Peio K.J., Snyder D.E., Duncan J.C., McFarren M.R. An advanced knowledge and design acquisition methodology for the pilot's associate //Harry C. Armstrong Aerospace Medical Research Laboratory, Human Systems Division, Tech. Rep., 1990.
5. Villalon Jorge, Calvo Rafael Analysis of a gold standard for concept map mining-how humans summarize text using concept maps //URL: https://cmc.ihmc.us/cmc2010papers/cmc2010-b4.pdf (дата обращения: 09.04.2023).
6. Trochim W.M.K. An introduction to concept mapping for planning and evaluation, Evaluation and Program Planning //vol. 12, no. 1, pp. 1–16, January 1989.
7. Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing //Cambridge University Press, 1999.
8. Das D., Matrins A.F.T. A survey on automatic text summarization //URL: http://www.cs.cmu.edu/~afm/Home_files/Das_Martins_survey_summarization.pdf (дата обращения 11.04.2023)
9. Spärck Jones K., Automatic summarising: The state of the art //Inf. Process. Manage., vol. 43, pp. 1449–1481, November 2007.
10. Binwahlan M.S., Salim N., Suanmali L. Fuzzy swarm-based text summarization //Journal of Computer Science, vol. 5, no. 5, pp. 338–346, 2009.
11. Spacy ru_news_lg model pipelines // URL: https://spacy.io/models/ru#ru_core_news_lg (дата обращения 11.04.2023)
12. Николаев И.С., Митренина О.В., Ландо Т.М. Прикладная и компьютерная лингвистика //Москва, Ленанд, 2017. ISBN: 978-5-9710-3472-8.
13. Wenxuan Zhou, Kevin Huang, Tengyu Ma, Jing Huang, Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling //URL: https://arxiv.org/pdf/2010.11304.pdf (дата обращения: 10.04.2023)
14. Hearst M.A. The debate on automated essay grading //IEEE Intelligent Systems and their Applications, vol. 15, no. 5, pp. 22-37, Sept.-Oct. 2000, doi: 10.1109/5254.889104.