Применение технологии интеллектуального анализа текста для решения задач управления проектами

А. А. Васильев
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
tolya051996@mail.ru

А. В. Горячев
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
avgoryachev@gmail.com

***Аннотация*. В данной статье анализируется возможность применения методов интеллектуального анализа текста для помощи в решении задач управления проектами. Интеллектуальный анализ текста – это область искусственного интеллекта, применение методов которой позволяет находить в текстовых массивах полезную, ранее неизвестную информацию. В связи с тем, что в ходе выполнения проектов генерируется много текстовых данных, представленных в различной форме (бумажные и электронные документы, письма, сообщения между участниками проектов и т. д.) и зачастую неструктурированных, необходимы технические средства, которые могли бы обрабатывать эту информацию и извлекать из нее полезные знания, которые могут помочь достичь целей проектов быстрее. В статье рассматривается сама технология интеллектуального анализа текста, основные подходы и методы, которые существуют в этой области на данный момент. Делаются предположения о том, как именно применение алгоритмов интеллектуального анализа текста может помочь в решении задач управления проектами.**

Ключевые слова: интеллектуальный анализ текста; Text Mining; управление проектами; искусственный интеллект

I. Введение

Современный человек живет и осуществляет свою деятельность в реальности, насыщенной непрерывно увеличивающимися потоками разнородной информации, хранящейся в различной форме во множестве источников. В частности, в связи с развитием современных технологий и их повсеместным использованием во всех сферах деятельности человека генерируется огромное количество цифровых данных в виде неструктурированного текста.

Если мы говорим о компаниях, ведущих множество проектов и занимающихся исследованиями или разработкой новых изделий, то большая часть корпоративной информации, примерно 80 %, доступна в текстовых форматах данных. Это может быть информация как по текущим, так и по завершенным проектам компании или конкретного отдела. Вместе с тем важно отметить, что зачастую руководителям, разработчикам и другим участникам проекта может быть полезно обратиться к информации, хранящейся по другим проектам – уже завершенным или идущим параллельно – с целью получения необходимой и полезной информации (такой информацией могут быть сведения о квалификации исполнителей и сроках выполнения проектов, о конкретных принятых технических решениях и реализованных «ноу-хау», бюджетах проектов или их этапов) для того, чтобы принимать в кратчайшие сроки максимально эффективные и результативные решения. Соответственно, встает вопрос реализации и практического применения инструментов, предназначенных для обработки текста на естественном, «человеческом» языке.

Целью данной статьи является исследование современного состояния теории, методов, задач и моделей в области обработки текста на естественном языке и интеллектуального анализа текста и выработки основных направлений применения достижений в этих областях для решения задач управления проектами.

II. Интеллектуальный анализ текста

Среди методов интеллектуального анализа данных (англ. «Data Mining», DM) наиболее популярной и актуальной остается технология интеллектуального анализа текста (англ. «Text Mining», TM). Она нацелена на извлечение значимых сведений из неструктурированных массивов текстовых документов. Данная технология предусматривает выявление в тексте шаблонных формулировок путем применения инструментов статистического изучения шаблонов.

Главной задачей TM является облегчение поиска важной информации в больших объемах текста посредством глубинного анализа, который заменяет традиционный ручной поиск, осуществляемый человеком.

Text Mining – это алгоритм автоматического определения в массивах «сырых» данных заданных показателей, неизвестных корреляций и знаний, которые несут в себе потенциальную ценность и могут использоваться человеком для дальнейшей работы или принятия решений. Так, сферами применения результатов интеллектуального анализа являются экономическая (анализ рынков), математическая или статистическая (прогнозирование), общественно-политическая (анализ социальной обстановки). Таким образом, можно говорить о высоком значении Text Mining в системе управления знаниями.

Подобные технологии незаменимы для извлечения знаний и играют немаловажную роль во всей системе управления знаниями.

При этом обнаруживаемая и извлекаемая информация должна быть «высококачественной». Данный термин предполагает, что получаемые данные достоверны, релевантны и полезны для дальнейшего исследования. Среди задач ТМ-технологии выделяются следующие: определение текстовых кластеров; формирование концептов; определение объектов; таксономизация; анализ смыслов; моделирование.

III. Этапы интеллектуального анализа текста

Выполнение технологической процедуры Text Mining разделено на пять этапов:

1) Поиск информации. На этом шаге осуществляется определение и подготовка исследуемой текстовой базы.

2) Предобработка текстов. Перед алгоритмической обработкой текстовые документы необходимо привести к соответствующему виду путем удаления лишних лексических единиц, не влияющих на итоговые результаты.

3) Извлечение требуемой информации. Здесь алгоритмам Text Mining задается понятийно-терминологический аппарат, на который следует опираться во время анализа.

4) Применение методов Text Mining. Основной шаг анализа, предусматривающий поиск и формирование новых знаний и взаимосвязей.

5) Анализ и интерпретация полученных результатов. Извлеченные данные организуются в выбранной исследователем форме, в том числе в виде графического или текстового документа.

Процесс предварительной текстовой обработки основывается на последовательном выполнении следующих приемов.

Предварительная обработка начинается с ***токенизации*** текста. Исходный текстовый массив разделяется на структурные элементы, называемые *токенами*. По своей форме они напоминают стандартные фрагменты текста – абзацы, предложения, отдельные слова.

Дальнейшая процедура обработки ориентируется на ***удаление стоп-слов***. К стоп-словам относят любые лексические единицы, не несущие смыслового содержания в рамках конкретного документа, а также все вспомогательные речевые единицы – междометия, союзы, предлоги, частицы, артикли и пр. Для оптимизации процесса изъятия перечень стоп-слов создается заблаговременно, основываясь на языке обрабатываемого текста.

Оставшиеся после удаления слова необходимо привести к стандартизированному образцу, возвратив лексемам их нормальный вид путем восстановления именительного падежа, единственного числа, орфографии по словарю. Данный прием называется ***стэмминг***, или ***лемматизация.*** Одним из результатов его реализации является нарушение семантических конструкций, что важно учитывать при определении языка текста. Стэмминг проводится программным способом на основе алгоритмов, наиболее распространенным из которых для русского языка является Snowball. Он подразумевает обнаружение в тексте однокоренных слов с последующим отсоединением от них словообразующих морфем.

Преобразование регистра слов. На данном этапе производится преобразование символов слов к одному регистру (верхнему или нижнему). Примером могут служить слова «text», «Text», «TEXT», приводимые к нижнему регистру «text».

Рис. 1. Этапы Text Mining

IV. Задачи интеллектуального анализа текста

Установленная методология Text Mining обеспечивает выполнение следующих основных задач:

1) классификация (classification) – включение текстовых документов в систему заранее заданных категорий;

2) кластеризация (clustering) – распределение всей совокупности документов на классы путем определения их семантических признаков с последующим объединением схожих документов в один класс;

3) построение семантических сетей – организация присутствующих в документе дескрипторов (ключевых фраз) на основе их семантической связи, обеспечивающая удобство навигации;

4) извлечение фактов, содержательных понятий (feature extraction) – установление в тексте фактов и отношений между ними;

5) аннотирование (summarization) – сокращение общего текстового объема без потери смысла;

6) индексирование тем (thematic indexing);

7) поиск по ключевым словам (keyword searching);

8) визуализация.

Рассмотрим чуть более подробно формулировки ***задач классификации*** и ***кластеризации***.

Задачу классификации определяют следующим способом. Пусть множество анализируемых документов

представлено в виде $D = \{d_1, \dots, d_i, \dots, d_n\}$, а множество категорий документов $C = \{c_1, \dots, c_r, \dots, c_m\}$. Присущее категориям множество признаком определяется следующим образом:

$$F(C) = \cup F(c_r),$$

где $F(c_r) = < t_1^r, \dots, t_k^r, \dots, t_z^r >$.

В результате анализируемый текст обзаводится т.н. *словарем*, то есть совокупностью лексических единиц, характеризующих все множество признаков документа, которые позволяют отнести его к тому или иному классу.

Вместе с тем, количество выявленных признаков текстового документа должно численно совпадать с множеством классовых признаков, высчитываемых по формуле $F(d_i) = < t_1^i, \dots, t_l^i, \dots, t_y^i >$. При этом признаки классов не должны противоречить и отличаться по количеству от множества признаков документов:

$$F(C) = F(D) = \cup F(d_i).$$

Таким образом, решение о принадлежности текста d_i к категории c_c принимается путем вычисления $F(d_i) \cap F(c_r)$.

Задачей кластеризации является разделение текстового массива на *кластеры* – группы однородных объектов, имеющих определенный набор признаков.

Дано конечное множество объектов $I = \{i_1, \dots, i_j, \dots i_n\}$. Каждый из объектов характеризуется m-компонентным признаковым описанием $(x_1, \dots, x_k, \dots, x_m), x_k \in X_k$, где X_k – допустимое множество значений признака. Требуется построить множество кластеров C и отображение $F: I \to C$. Кластер $c_h \in C$ имеет структуру $c_h = \{i_j, i_p: i_j, i_p \in I, d(i_j, i_p) < \sigma\}$, т.е. кластер состоит из объектов, находящихся в пространстве признаков рядом в смысле метрики d и определяется величиной σ.

Для алгоритмической кластеризации необходимо привести текстовую информацию к формату т.н. модели векторного пространства Vector Space Model. Она представляет собой графический метод отображения семантической схожести объектов. Благодаря простоте использования и наглядности визуализации данная модель получила широкое распространение.

Основой Vector Space Model является многомерное пространство, где каждое измерение представляет собой слово из выбранного текстового набора. В результате создается матрица из слов и документов:

$$M = | F | \times | D |,$$

где $F = \{f_1, \dots, f_k, \dots f_z\}; D = \{d_1, \dots, d_i, \dots, d_n\}, d_i$ – вектор в z-мерном пространстве R^z.

Каждому признаку f_k в документе d_i ставится в соответствие его вес $\omega_{k,i}$, который обозначает важность этого признака для данного документа.

V. Применение интеллектуального анализа текста для решения задач управления проектами

Прежде чем говорить о применении технологии Text Mining для решения задач в области управления проектами, необходимо отметить несколько специфических особенностей сферы управления проектами.

Во-первых, в ходе работы над проектами генерируется большой объем текстовой информации, представленный в различной форме – письма, файлы с документацией, планы по управлению и распределению ресурсов, инструкции, поручения, приказы и т. д. Все эти данные и файлы имеют различные форматы – текстовые документы, документы Microsoft Word, XML-документы, прочие специфические форматы. Поэтому первой задачей, которую необходимо решить, является разработка инструментов для извлечения текста с полезной информацией для управления проектной деятельностью из данных файлов.

Во-вторых, несмотря на то, что область управления проектами достаточно универсальна, все же необходимо отметить, что в массивах текстовой информации по разработке того или иного инженерного изделия встречается большое количество технических терминов, дат (включая сроки выполнения задач и этапов проекта), имен исполнителей и заказчиков, спецификаций материалов и инструментов. Поэтому требуется разработка новых (или модификация существующих) эффективных методов для извлечения данных терминов, дат, имен для их последующего анализа и, возможно, даже построения графов связей с целью их последующего анализа и повышения качества и скорости работы над проектами. Также нельзя не отметить, что зачастую в среде разработчиков и конструкторов присутствуют разногласия в трактовке тех или иных технических терминов – а потому перед этапом извлечения их из текстов может потребоваться, в том числе, и разработка онтологии (единого пространства понятий) конкретной предметной области того или иного проекта.

В-третьих, требуется разработка новых (или применение уже существующих) математических методов и метрик оценки качества выполнения процессов Text Mining на всех этапах – начиная от извлечения текстовой информации из документов различного формата и заканчивая выполнением поисковых запросов пользователями-участниками проектной деятельности. Это позволит иметь ясную картину того, насколько эффективно работают выбранные для реализации решения – и какой вклад они вносят в повышение эффективности управления проектами внутри организации.

С учетом всего вышесказанного, авторами предложен для будущей реализации следующий функционал предполагаемой программной системы, анализирующей, хранящей и представляющей информацию о проектах в организациях:

1) разбор структуры файлов различных форматов, связанных с выполнением проектов и содержащих текстовую информацию, с эффективным и максимально полным извлечением последней с применением алгоритмов обработки текстов на естественном языке;

2) обеспечение эффективного хранения, доступа и поиска информации как внутри исходных файлов и данных, так и внутри данных, полученных после извлечения текста и его анализа;

3) извлечение метаданных из исходной текстовой информации, содержащих в сжатом виде основную сводку об исходном документе или тексте;

4) извлечение, анализ и хранение упоминаемых в документах имен людей/организаций (исполнителей, заказчиков, назначенных ответственных), а также дат и сроков выполнения как всего проекта, так и отдельных его этапов и задач с последующим представлением в том или ином визуальном виде и/или генерацией отчетов по текущему состоянию;

5) построение на основе извлеченных имен людей/организаций связей между ними, тенденций и отношений для наиболее эффективного и точного планирования и управления ресурсами (включая человеческие);

6) обеспечение классификации, категоризации, эффективного хранения и доступа к текстовой информации на основе ее содержимого;

7) автоматическое аннотирование и генерация сводных документов и отчетов, в которых в краткой форме будет изложено содержание целых групп отчетов;

8) обеспечение непрерывной оценки качества работы всех модулей программной системы с целью улучшения применяемых математических методов и метрик.

VI. Заключение

Интеллектуальный анализ текстовых данных становится все более важным научным методом в последние годы, и его потенциал очень высок.

С одной стороны, большая часть информации находится в текстовом виде, а с другой – полезные знания и достоверная информация играют очень важную роль в успехе бизнеса и организации. Автоматизированный сбор информации и связанное с этим появление огромных объемов данных делают автоматический анализ данных необходимым во многих областях.

В частности, неуклонный рост текстовых данных в проектных организациях (бумажные и электронные документы, письма, сообщения между участниками проектов и т. д.) делает все более важными компьютерные методы классификации, кластеризации, NLP, извлечения и поиска информации. Можно прогнозировать, что в будущем наибольший успеха смогут добиться компании, которые смогут максимально эффективно использовать абсолютно весь объем накопленных ими данных о проектах – а, следовательно, производить эффективный подбор исполнителей, распределение задач между ними, производить точный расчет сроков и бюджетов проектов, предусматривать и на самых ранних этапах минимизировать риски, основываясь на ранее полученных и хранящихся знаниях.

Таким образом, применение и дальнейшее развитие методов и приемов Text Mining позволит в долгосрочной перспективе сократить расходы организаций на ведение проектов, а также будет способствовать повышению эффективности и прибыльности их работы

Список литературы

[1] Барсегян А.А., Куприянов М.С., Холод И.И., Тесс М.Д., Елизаров С.И. Анализ данных и процессов: учеб. пособие. СПб.: БХВ-Петербург, 2009. 512 с.

[2] Молнина Е.В., Картуков М.С. Проблема интеллектуального анализа текстов // Фундаментальные исследования. 2007. №11. С. 136-137.

[3] Певченко С.С. Методы интеллектуального анализа данных // Молодой ученый. 2015. №13 (93). С. 167-169.

[4] George, G., Osinega, E.C., Lavie, D., and Scott, B.A. (2016) From the editors – big data and data science methods for management research. Academy of Management Journal, 59, 5, 1493–1507.

[5] Ghazinoory, S., Ameri, F., and Farnoodi, S. (2013) An application of the text mining approach to select technology centers of excellence. Technological Forecasting and Social Change, 80, 5, 918–931.

[6] Goffin, K., Ahlström, P., Bianchi, M. and Richtnér, A. (2019) Perpsective: state-of-the-art: the quality of case study research in innovation management. Journal of Product Innovation Management, 36, 5, 586–615.

[7] Hoornaert, S., Ballings, M., Malthouse, E.C., and Van den Poel, D. (2017) Identifying new product ideas: waiting for the wisdom of the crowd or screening ideas in real time. Journal of Product Innovation Management, 34, 5, 580–597.

[8] Janasik, N., Honkela, T., and Bruun, H. (2009) Text mining in qualitative research: application of an unsupervised learning method. Organizational Research Methods, 12, 3, 436–460.