В перспективе вызывает интерес выявление частот, для которых наиболее явно прослеживаются влияние свойств отражающей поверхности, а также оценка влияния запыленности и влажности, так как эти факторы могут оказаться не менее важными в части обеспечения работы периферийных устройств ЭВМ при контроле состояния и наличия расходных материалов.

Литература

1. Ультразвуковой датчик измерения расстояния HC-SR04 // Автоматика и программная инженерия. 2017. № 4 (22). URL: http://www.jurnal.nips.ru

2. Руководство пользователя HANTEK 6000BC/BD/6074BE [электронный ресурс]. URL: http://www.hantek.ru/products/mans/HT6004BX_rus.pdf

Literatura

1. Ultrazvukovoj datchik izmereniya rasstoyaniya HC-SR04 // Avtomatika i programmnaya inzheneriya. 2017. № 4 (22). URL: http://www.jurnal.nips.ru

2. Rukovodstvo polzovatelya HANTEK 6000BC/BD/6074BE [elektronnyj resurs]. URL: http://www.hantek.ru/products/mans/HT6004BX_rus.pdf

---

УДК 004.82

О.В. Золотарев, В.А. Шуйнов, С.М. Крестьянинов, Д.А. Соловьев,
Российский новый университет

МЕТОДЫ ВЫДЕЛЕНИЯ СУЩНОСТЕЙ, ИХ АТРИБУТОВ, ПРОЦЕССОВ ИЗ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА И ПОСТРОЕНИЕ СЕМАНТИЧЕСКОЙ СЕТИ

Рассматриваются способы извлечения объектов, их характеристик, связанных с ними действий из текстов естественного языка с последующим выделением биз-

нес-процессов, характерных для конкретной предметной области и формированием тезаурусов понятий, действий, бизнес-процессов по конкретным предметным областям. Это позволит в дальнейшем при выделении из текстов естественного языка бизнес-процессов для определенных предметных областей, для которых уже были построены тезаурусы, согласовывать вновь построенные элементы с представленными в тезаурусах объектами и процессами. Описываются методы первичной классификации текстов при соотнесении их с конкретной предметной областью, позволяющие в существенной степени сократить количество нерелевантных документов.

Ключевые слова: знания, семантические сети, фрагменты знаний, объекты, процессы, тезаурусы.

O.V. Zolotarev, V.A. Shuinov, S.M. Krestyaninov, D.A. Soloviev,
Russian New University

METHODS FOR ISOLATION OF ESSENCES, THEIR ATTRIBUTES, PROCESSES FROM TEXTS OF NATURAL LANGUAGE AND CONSTRUCTION OF A SEMANTIC NETWORK

The article discusses methods of extracting objects, their characteristics, related actions from natural language texts, followed by the selection of business processes characteristic of a specific subject area (software) and the formation of thesauri of concepts, actions, business processes in specific subject areas. This will allow in the future, when extracting business processes from natural language texts for certain subject areas for which thesauri have already been built, to coordinate the newly constructed elements with the objects and processes presented in thesauri. The article describes the methods of primary classification of texts when correlating them with a specific subject area, which can significantly reduce the number of irrelevant documents.

Keywords: knowledge, semantic networks, fragments of knowledge, objects, processes, thesaurus.

Введение

Для отработки методов построения моделей бизнес-процессов в сфере деловой прозы предполагается детальное изучение предметной области, исследование стандартов моделирования, используемых как в проектах внедрения корпоративных информационных систем (ИС), так и в проектах по повышению эффективности деятельности предприятий. В процессе исследования деятельности предприятия строятся динамические и статические модели предметной области [6]. В данной работе описываются методы выделения объектов и процессов из текстов естественного языка на основе анализа документов большого объема, так называемых Big Data. Данный подход основывается на построении и ведении тезаурусов объектов и процессов.

Состояние вопроса

Себастьян Падо из Саарского университета (Германия) и Мирелла Лапата из Шеффилдского университета (Великобритания) рассматривают вопрос построения семантических пространств на основе традиционных векторных моделей с учетом синтаксических отношений [1]. Семантические свойства слов представляются в виде частотной матрицы, каждый ряд которой соответствует уникальному целевому слову, а каждая колонка – лингвистическому контексту. Семантическая информация извлекается из текстов большого объема на основе анализа окружения слова. Слово рассматривается как точка в многомерном семантическом пространстве. На основе близости между точками семантического пространства вычисляется семантическое сходство между словами с использованием метрик. Анализ семантического сходства выполняется на основе статистических методов с расчетом частотности появления в тексте близких точек семантического пространства. В процессе

анализа рассматриваются метрики Euclidean, Jaccard's, Kullback-Leibler и другие [4]. По результатам исследования делается вывод о том, что контекстное окружение играет важную роль в распознавании лексических отношений между словами.

У Э. Путрич и А.В. Карк (2007, 2008) выделяются бизнес-правила на основе анализа естественно-языковых текстов в форме <Условие><Действие> [3].

В системе Rocket AeroText информация извлекается из текстов естественного языка на основе анализа больших объемов текстов с целью выделения бизнес-процессов [2].

Методы обработки текстов

При обработке текстов естественного языка могут использоваться многочисленные инструменты для синтаксического и семантического анализа. В рамках данной работы был использован оригинальный инструмент Pullenti, разработанный в Институте проблем информатики [7]. Этот инструмент позволяет проводить не только морфологический, но и синтаксический анализ текстовой информации, а также частично семантический анализ [8]. На основе выделенной из текстов информации выделяются ключевые слова, частота их появления, кластеры, строится семантическое пространство (СП). Значимость ключевых слов определяется с учетом частотности появления термина и его окрестности:

<Ключевое слово, Кластер, Значимость> (1)

Окончательное решение о включении ключевого слова в семантическое пространство принимает группа экспертов.

Рассчитывается вероятность правильного решения i-го эксперта (эксперт принимает решение о принадлежности данного документа к определенной категории):

$$P_i = p\,(\text{Кластер} \mid \text{Ключевое слово}), \qquad (2)$$

где P_i – вероятность корректного решения i-го эксперта.

Значимость мнений экспертов вычисляется по формуле:

$$W_i = \log(P_i / 1 - P_i). \qquad (3)$$

Одной из важнейших задач при формировании базы знаний является функция обобщения, которая позволяет не только достраивать базу знаний, но и создавать метазнания системы, необходимые как для построения моделей бизнес-процессов, так и для их коррекции [5].

При формирования семантического пространства в соответствии с требованиями бизнес-процессов конкретной предметной области [8] выделяются элементы, описывающие:

– объекты (сущности),
– атрибуты объектов,
– процессы;
– связи между объектами;
– входную и выходную информацию для процессов;
– регламентную информацию о выполнении процесса;
– используемые при выполнении процессов ресурсы;
– используемые в процессе механизмы и технологии;
– потоки данных между процессами;
– общее описание процесса.

Для построения семантической сети первоначально фрагмент текста естественного языка обрабатывается инструментом Pullenti, затем с помощью специального программного обеспечения производится анализ структуры текста.

При этом выбрана часть текста, которая является минимальной и достаточной для формирования представленных фрагментов семантической сети, описываю-

щих процесс изготовления детали из заготовки: «Деталь изготавливается из заготовки на основе чертежа. Деталь изготавливает токарь. Заготовка подготавливается из бруса. Из бруса вытачивается деталь».

В результате обработки данного текста строятся следующие фрагменты семантической сети:

Изготовление (Процесс, _) (4)
Заготовка (Объект, Вход, Изготовление (5)
Деталь (Объект, Выход, Изготовление) (6)
Чертеж (Объект, Управление, Изготовление) (7)
Токарь (Объект, Участник, Изготовление) (8)

Для построения семантической сети использована логика предикатов первого порядка.

Данный подход позволяет не только строить семантическое пространство на основе анализа текстов, но и попутно формировать тезаурусы предметный областей. Автоматическое построение тезаурусов и баз знаний предполагает наличие определенного количества неточностей и ошибок при формировании базы знаний и тезаурусов. Поэтому на первоначальном этапе формирования знаний по предметным областям допускается наличие ручного труда в значительной степени. Но в процессе обработки представительного корпуса текстов естественного языка, уточнения и расширения базы знаний количество ручного труда будет сокращаться в существенной степени. Это позволит практически в автоматическом режиме формировать модели бизнес-процессов предметной области на основе накопленных знаний.

Представленные в статье подходы к анализу текстов естественного языка и формированию семантического пространства позволяют не только выделить объекты, их атрибуты, процессы, связи между объектами, но и сформировать тезаурусы предметной области.

На основе сформированной семантической сети (семантического пространства) могут быть построены модели предметной области в различных нотациях и погружены в различные информационные среды. Это может в существенной степени облегчить труд аналитиков на этапе обследования предприятия, а также оказать существенную помощь при проектировании структуры информационной системы.

Литература

1. *Putrycz E., Kark A.W.* Recovering Business Rules from Legacy Source Code for System Modernization. Published in RuleML 2007. Computer Science.

2. Rocket AeroText. BI Tools. URL: softwareconnect.com

3. *Padó S., Lapata M.* Dependency-based construction of semantic space models. Computational Linguistics. MIT Press. Vol. 33. № 2. P. 161–199.

4. *Cunha V., Zavala A., Inácio P.R.M., Magoni D.* Classification of Encrypted Internet Traffic Using Kullback-Leibler Divergence and Euclidean Distance // Advanced Information Networking and Applications. March, 2020. P. 883–897.

5. *Золотарев О.В., Козеренко Е.Б., Шарнин М.М.* Принципы построения моделей бизнес-процессов предметной области на основе обработки текстов естественного языка // Вестник Российского нового университета. Сер.: Сложные системы: модели, анализ и управление. 2014. № 4. С. 82–88.

6. *Золотарев О.В.* Формализация знаний о предметной области на основе анализа естественно-языковых структур // Цивилизация знаний: проблема человека в науке XXI века: тр. XII Междунар. науч. конф. 2011. С. 78–80.

7. *Золотарев О.В., Шарнин М.М., Клименко С.В., Кузнецов К.И.* Система PullEnty – извлечение информации из текстов естественного языка и автоматизированное построение информационных систем // Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности (SCVRT2015-16): тр. Междунар. науч. конф.: в 2 т. 2016. С. 28–35.

8. *Золотарев О.В, Шарнин М.М.* Методы извлечения знаний из текстов естественного языка и построение моделей бизнес-процессов на основе выделения процессов, объектов, их связей и характеристик // Междунар. науч. конф. Моск. физико-технич. института (гос. ун-та) Ин-та физико-технич. информатики. Институт физико-технической информатики, 2015. С. 92–98.

Literatura

1. *Putrycz E., Kark A.W.* Recovering Business Rules from Legacy Source Code for System Modernization. Published in RuleML 2007. Computer Science.

2. Rocket AeroText. BI Tools. URL: softwareconnect.com

3. *Padó S., Lapata M.* Dependency-based construction of semantic space models. Computational Linguistics. MIT Press. Vol. 33. № 2. P. 161–199.

4. *Cunha V., Zavala A., Inácio P.R.M., Magoni D.* Classification of Encrypted Internet Traffic Using Kullback-Leibler Divergence and Euclidean Distance // Advanced Information Networking and Applications. March, 2020. P. 883–897.

5. *Zolotarev O.V., Kozerenko E.B., Sharnin M.M.* Principy postroeniya modelej biznes-processov predmetnoj oblasti na osnove obrabotki tekstov estestvennogo yazyka // Vestnik Rossijskogo novogo universiteta. Ser.: Slozhnye sistemy: modeli, analiz i upravlenie. 2014. № 4. S. 82–88.

6. *Zolotarev O.V.* Formalizaciya znanij o predmetnoj oblasti na osnove analiza estestvenno-yazykovyx struktur // Civilizaciya znanij: problema cheloveka v nauke XXI veka: tr. XII Mezhdunar. nauch. konf. 2011. S. 78–80.

7. *Zolotaryov O.V., Sharnin M.M., Klimenko S.V., Kuzneczov K.I.* Sistema PullEnty – izvlechenie informacii iz tekstov estestvennogo yazyka i avtomatizirovannoe postroenie informacionnyx sistem // Situacionnye centry i informacionno-analiticheskie sistemy klassa 4i dlya zadach monitoringa i bezopasnosti (SCVRT2015-16): tr. Mezhdunar. nauch. konf.: v 2 t. 2016. S. 28–35.

8. *Zolotaryov O.V, Sharnin M.M.* Metody izvlecheniya znanij iz tekstov estestvennogo yazyka i postroenie modelej biznes-processov na osnove vydeleniya processov, obektov, ix svyazej i xarakteristik // Mezhdunar. nauch. konf. Mosk. fiziko-texnich. instituta (gos. un-ta) In-ta fiziko-texnich. informatiki. Institut fiziko-texnicheskoj informatiki, 2015. S. 92–98.