

Список литературы

1. Зенина О.В. Информационное общество: цифровое настоящее и цифровое будущее, достижения и противоречия / О.В. Зенина // Электронный журнал «Знание. Понимание. Умение». – 2009. – № 4.
2. Хаперская А.В. / Возникновение необходимости создания нового экономического механизма в связи с появлением программ корпоративной социальной ответственности / А.В. Хаперская, С.З. Мусина // Современные проблемы науки и образования. – 2015. – № 1.
3. Голобуцкий А., Шевчук О. Электронное правительство. – М.: Знание, 2007.
4. Ирхин Ю.В. «Электронное правительство»: теория и практика // Гос. служба. – 2008. – № 4.
5. Попова Л.Л., Кирекова О.А. Концепция электронного правительства: современное состояние // Трансформация научных парадигм и коммуникативные практики в информационном социуме: сборник научных трудов конференции, Томск, 22–23 Ноября 2012. – Томск: ТПУ, 2013 – С. 296–298.

ИДЕНТИФИКАЦИЯ ОСТАНОВОК ГОРОДСКОГО ОБЩЕСТВЕННОГО ТРАНСПОРТА ПОСРЕДСТВОМ СИНТАКСИЧЕСКОГО АНАЛИЗА ИХ ТЕКСТОВЫХ ОПИСАНИЙ

А.З. Чинь, В.Ю. Погребной

(г. Томск, Томский политехнический университет)

E-mail: tad3012@gmail.com, vadimshah.1990@gmail.com

URBAN PUBLIC TRANSPORT STOP IDENTIFICATION USING SYNTACTIC ANALYSIS ON TEXT DESCRIPTION

A.D. Trinh, V.Y. Pogrebnoy

(Tomsk, Tomsk Polytechnic University)

Nowadays, with the use of information technologies the quality of the urban public transportation can be noticeably increased. It is important to understand local citizens' demands and provide good information services. This article deals with the problem of identifying bus stops using their text descriptions received from users. The issue consists mainly of the difficulties in language understanding between machine and human and syntactically analyzing text messages.

Key words: Public transport, syntactic analysis, information technology, natural language processing, text message, algorithm, minimum distance.

Введение. Во многих городах мира задача качественной организации и управления транспортным движением является актуальной нерешенной задачей. Актуальность данного вопроса заключается в необходимости применения современных технологий и оборудования, развития системы общественного транспорта, повышения качества обслуживания жителей и предоставления информационных услуг с использованием интеллектуальных систем. Решение данной задачи требует совместного применения достижений в разных областях науки, в том числе в математике, информационных системах и технологиях, социальной науке.

В транспортной системе города Томска существует множество проблем, требующих немедленного решения, в частности существует необходимость повышения качества услуг пассажирских перевозок общественным транспортом. Среди инструментов, способствующих решению данной проблемы, одним из самых эффективных является информационная система, предоставляющая различные информационные услуги в режиме реального времени. Пользователи информационной системы могут получать различную актуальную информацию: расписание движения транспорта, прогноз времени прибытия транспортных

средств, кратчайшие маршруты движения и др. Для реализации системы, способной предоставлять услуги по запросу пользователя, сформированного в произвольной форме, была поставлена задача разработки модуля идентификации по предоставленному текстовому описанию различных объектов городской улично-дорожной сети и, в частности, остановок общественного транспорта.

Методология. Решение задачи идентификации объектов посредством синтаксического анализа текстовых описаний требует изучения и применения знаний раздела науки, называемого обработкой естественного языка. Обработка естественного языка (ОЕЯ, англ. язык Natural language processing, NLP) представляет собой общее направление компьютерной науки, искусственного интеллекта и математической лингвистики, изучающее проблемы компьютерного анализа и синтеза естественных языков. ОЕЯ относится к области человеко-машинного взаимодействия [0].

В ОЕЯ одним из часто используемых алгоритмов является дерево принятия решений. Дерево принятия решений – это средство поддержки принятия решений, использующееся в статистике и анализе данных. Структура дерева представляет собой «листья» и «ветки». На ребрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция. В «листьях» записаны значения целевой функции, а в остальных узлах – атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Подобные деревья решений широко используются в интеллектуальном анализе данных. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе [2].

В процессе обработки текстов часто сталкиваются с проблемой преобразования одной строки в другую. Оптимальным решением данной проблемы называется расстояние между строками, которое рассчитывается как наименьшее количество шагов, необходимых для преобразования строк методом удаления, добавления или замены символов. Пусть имеются две строки: X длиной N и Y длиной M . Для нахождения расстояния между строками определим двумерный массив $D_{N,M}$ следующим образом:

- $D_{i,j}$ – минимальное расстояние между $X_{1..i}$ и $Y_{1..j}$;
- Для каждого $i = 1..N, j = 1..M$

$$\begin{aligned} D_{i,0} &= i \\ D_{0,j} &= j \\ D_{i,j} &= \min \begin{cases} D_{i-1,j} + 1 \\ D_{i,j-1} + 1 \\ D_{i-1,j-1} + \begin{cases} 2, \text{ если } X_i \neq Y_j \\ 0, \text{ если } X_i = Y_j \end{cases} \end{cases} \end{aligned}$$

В результате наименьшее количество шагов преобразования строки X в строку Y – это значение $D_{N,M}$. Результат преобразования можно отследить следующим образом:

- если $D_{i,j} = D_{i-1,j} + 1$, то нужно удалить символ X_i ;
- если $D_{i,j} = D_{i,j-1} + 1$, то нужно добавить символ Y_j ;
- если $D_{i,j} = D_{i-1,j-1} + 2$, то нужно заменить символ X_i на Y_j ;
- если $D_{i,j} = D_{i-1,j-1}$, то преобразование не требуется.

Алгоритм идентификации. С учетом особенностей русского языка и человеческой логики для обработки текстовых описаний был реализован алгоритм синтаксического анализа и схема сравнения слов.

Схема сравнения слова $W1$ со словом $W2$ может быть представлена в виде дерева принятия решений как на Рис. 11.

На каждом этапе работы разработанной схемы должно быть выполнено сравнение входного слова с множеством слов системы, называемым словарём. Пусть имеется набор слов: *Новосибирская*, *Сибирская*, *Новособорная*, *СибГМУ*. По схеме сравнения слов результатом поиска для слова «*Сибирская*» является только слово «*Сибирская*», так как слова полностью совпадают. Для входного текста «*Сиб*» будет найдено два слова «*Сибирская*» и «*СибГМУ*» на втором этапе схемы. Если на первых трех этапах в словаре не найдено ни одного подходящего слова, то выполняется поиск слов с минимальным расстоянием от входного слова.

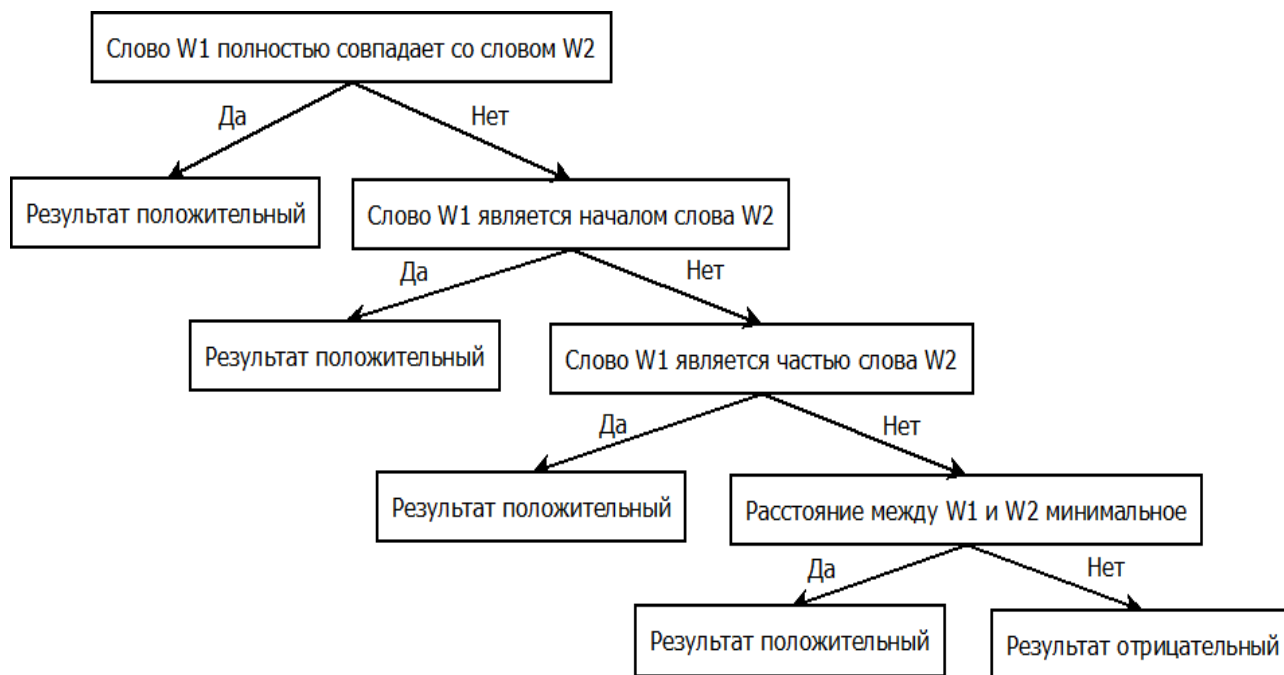


Рис. 11. Дерево принятия решений для сравнения двух слов

Каждая остановка описывается с использованием следующей информации: код, название, список альтернативных названий, описание и название улицы, на которой она находится. Данную информацию можно использовать в качестве словаря для поиска требуемой остановки. Алгоритм поиска остановки начинается с разбиения текстов предоставленного пользователем описания и имеющейся информации об остановках на отдельные слова и удаления вспомогательных слов и знаков препинания. Такое преобразование необходимо для дальнейшего сравнения отдельных слов вместо словосочетаний и предложений [3].

Алгоритм синтаксического анализа представлен на Рис. 12.

Согласно алгоритму процесс обработки входного текста проходит следующие этапы:

1. Происходит проверка наличия слова, начинающегося на «трам», во входном тексте. Если такое слово существует, то необходимо оформить список, состоящий только из трамвайных остановок. Иначе список состоит из всех остановок.

2. Для каждого слова входного текста согласно схеме, представленной на Рис. 11, выполняется сравнение с данными остановок: кодами, названиями, альтернативными названиями, описаниями, названиями улиц. При этом схема может изменяться в зависимости от порядкового номера обрабатываемого слова, наличия среди предыдущих слов специальных слов и др. (например, первым словом входного текста обязательно является часть названия остановки). Если совпадения с остановкой отсутствуют, то остановка удаляется из списка, так как она не имеет отношения к входным данным.

3. Алгоритм заканчивается в двух случаях: на каком-либо этапе в списке осталась одна остановка, или все слова во входном тексте были рассмотрены.

Результатом поиска являются остановки, оставшиеся в списке.

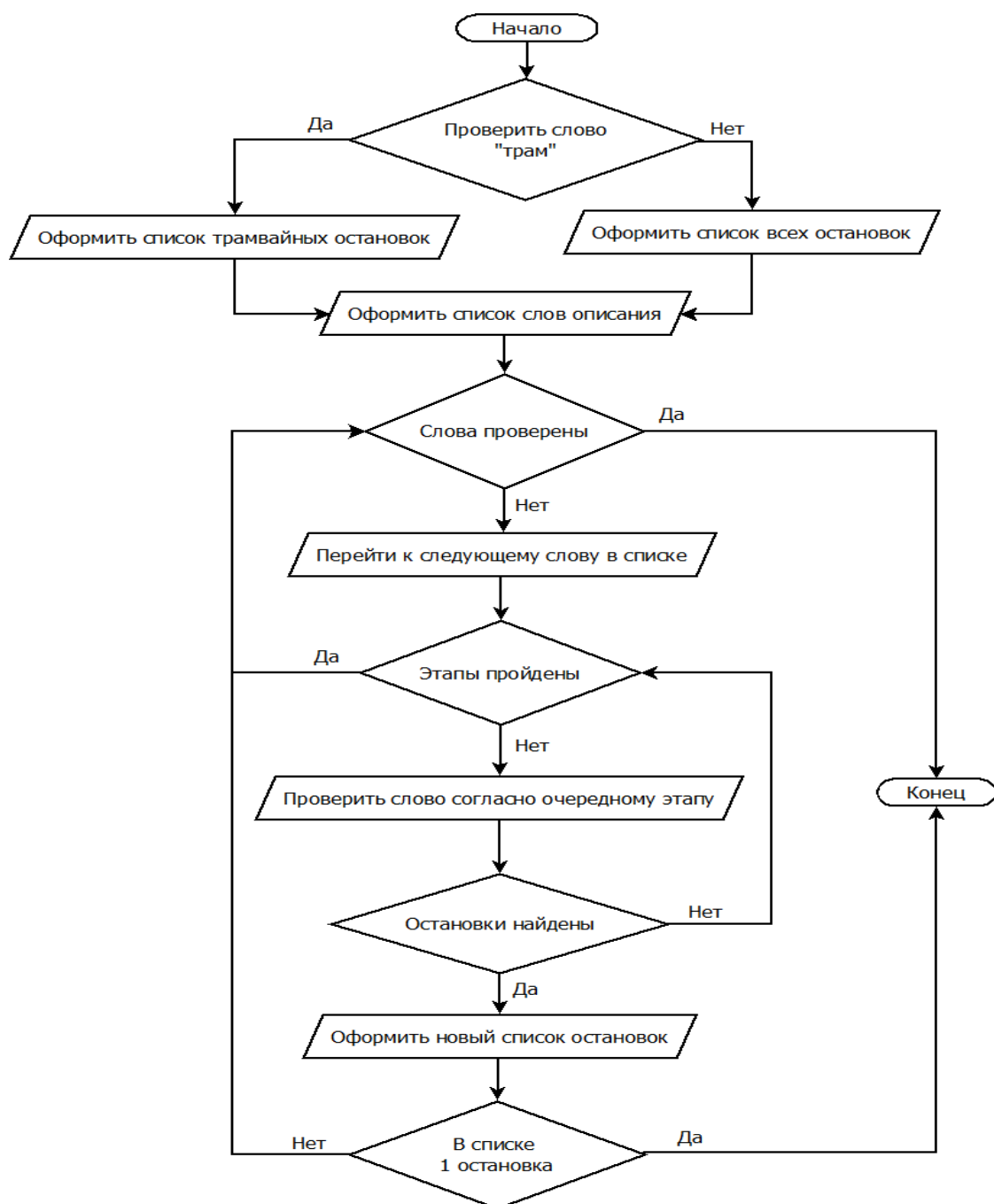


Рис. 12. Алгоритм обработки входного текста

Заключение. Результаты работы помогают частично решить вопрос повышения эффективности работы городской транспортной системы и качества обслуживания жителей общественным транспортом. Так, описанные методы и алгоритмы идентификации используются для определения остановок в маршрутной сети города Томска по описаниям, предоставляемым в смс-сообщениях. Пассажир, произвольно описав остановку, из которой он планирует движение, в ответ получает прогноз времени прибытия транспорта. Разработанная система позволяет сократить время ожидания транспорта на остановках, что указывает на практическую значимость результатов работы для населения города.

Список литературы

1. Christopher D. Manning, Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999. Электронный ресурс. URL: <http://cognet.mit.edu/library/books/view?isbn=0262133601> (Дата обращения 15.03.2015 г.)

2. Пескова О.В. Алгоритмы классификации полнотекстовых документов // Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. – М.: МИЭМ (Московский государственный институт электроники и математики), 2011. – С. 170–212.

3. Робин Хантер Основные концепции компиляторов = The Essence of Compilers. – М.: «Вильямс», 2002. – С. 256.

КУЛЬТУРНЫЕ АСПЕКТЫ СОЦИАЛЬНЫХ СЕТЕЙ

A.C. Чирва

(г. Томск, Томский государственный университет)

E-mail: loya-2345@mail.ru

CULTURAL ASPECTS OF SOCIAL NETWORKS

A.S. Chirva

(Tomsk, Tomsk State University)

In my article I raised the topic of social networks, as they appeared in our lives. What has changed since their appearance as they affected the lives of mankind.

Key words: social networks, culture, communication, society, world, people.

Общие данные об интернете и социальных сетях. Интернет в наше время является неотъемлемой частью современной культуры. В глубинах интернета появились социальные сети, зарождение социальных сетей происходило очень быстрыми темпами. Социальные сети показали себя в роли нового и устойчивого феномена культуры и с увеличением популярности в обществе на современном этапе они влияют на него все больше и больше. Анализируя статистические данные, можно отметить, что сейчас к социальным сетям подключены 2 млрд человек, а это, к слову, практически чуть меньше 1/3 населения мира. Наиболее популярной социальной сетью является Facebook, число его постоянных пользователей более 750 млн чел. Очень популярна социальная сеть ВКонтакте, которая стала известна за не большое количество времени, и на данный период выделяется наибольшим спросом аудитории пользователей [1].

Итак, в чем же заключается привлекательность социальных сетей для такого количества людей? На мой взгляд, это – соединение коммуникации и развлечений.

Необходимые составляющие социальных сетей. Социальные сети дают возможность превратить своего собеседника в некое средство развлечения. В таком общении нет ненужных эмоций, что позволяет людям защититься от лишних мелочей. Кроме того, следует помнить о важной составляющей социальных сетей: информация, расположенная в интернете, передается от одного пользователя к другому, благодаря этому зарождаются новые интересы, развлечения, то, что как раз и привлекает современного человека.

Термин «социальная сеть» можно трактовать как определенный круг знакомых человека и социальных связей между этими людьми.

Социальные сети в мире. По данным исследования одной известной международной компании, средний интернет-пользователь нашей страны в социальных сетях проводит 6,6 часов в месяц. Можно сказать, что помимо людей, совершенно не использующих социальные сети, существуют люди, которые способны там находиться все вечера, выходные и любое свободное время. И чем больше функций предлагает социальная сеть, тем больше времени готов посвящать ей человек.

В мире сейчас насчитывается около двадцати крупных социальных сетей, имеющих разный рейтинг в различных странах мира.