Раздел 4. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ЦИФРОВАЯ ЭКОНОМИКА

УДК 004.891.3, 004.93'14

РАЗРАБОТКА КЛАССИФИКАЦИОННОГО АЛГОРИТМА ТЕКСТОВОЙ ИНФОРМАЦИИ НА ОСНОВЕ ОБРАЩЕНИЙ ПОЛЬЗОВАТЕЛЕЙ ИНТЕРНЕТ-РЕСУРСОВ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Бобков Сергей Петрович (bsp@isuct.ru)
ФГБОУ ВО «Ивановский государственный химико-технологический университет»
Суворов Станислав Вадимович
Фролов Игорь Алексеевич
Казадаев Артем Ильич
ФГБОУ ВО «Московский политехнический университет»

В статье описываются возможности использования интеллектуальных систем для управления потоком сообщений в организации. Рассмотрен метод классификации текстовых обращений пользователей Интернет-ресурсов, базирующийся на анализе как смысловой, так и эмоциональной (тональности текста) составляющих обращений. Это является основным преимуществом по сравнению с имеющимися методиками классификации текстов. Эмоциональная составляющая определялась по частотным диспропорциям гласных букв в тексте, что позволило с высокой степенью надежности отделять малоинформативные обращения пользователей от реальных технически сложных заявок, и, таким образом, снизить нагрузку на службы технической поддержки Интернет-ресурсов. Для классификации использовалась комбинация нейронных сетей двух типов. На примере предприятия ООО «Лаборатория нейросетевых технологий» авторами работы была произведена апробация описанного метода, подтвердившая его эффективность.

Ключевые слова: нейронные сети, классификация текстов, тональность текста, «мешок» слов.

Введение. Служба технической поддержки любого интернет-ресурса сталкивается с большим объемом обращений пользователей этого ресурса. Среди этого потока сообщений имеются претензии финансового характера, описание проблем технического характера, а также большое количество сообщений, которые содержат негативные эмоции пользователей, не сумевших решить какие-либо свои проблемы.

В связи с этим очень важно правильно классифицировать сообщения пользователей и перенаправить их профильным специалистам: в бухгалтерию, юридический отдел, техническим специалистам и т.д.

Одним из вариантов решения данной задачи является введение промежуточной диспетчеризации сообщений пользователей (ее еще называют первой линией поддержки пользователей), как ручной, так и автоматизированной.

Ручная диспетчеризация, несмотря на свою эффективность, связана с определенными издержками экономического характера: организация рабочего места для сотрудника, охрана труда, выплата заработной платы и т.д. Именно поэтому внедрение автоматизированной системы диспетчеризации в службе поддержки интернет-ресурса выглядит экономически выгодным и привлекательным шагом.

Для корректной работы такой системы необходимо уметь классифицировать текстовую информацию, содержащуюся в обращениях пользователей, по заданным критериям.

В данной работе предлагается использовать искусственные нейронные сети [4, 7] в качестве классификаторов, причем алгоритм классификации должен учитывать не только смысловую составляющую обращения, но и его эмоциональный фон, поскольку именно на основании его обычно выставляется приоритет работы с обращением и сроки выполнения.

Таким образом, основная задача, которую решали авторы работы, была разработка методики представления текста в виде удобном для использования нейронными сетями и учитывающим как смысловую составляющую текста, так и его эмоциональную составляющую.

Описание алгоритма. В основу оценки эмоционального фона сообщений авторами была положена идея, используемая для выявления троллей в социальных сетях [6]. В указанной работе для выявления эмоционального состояния участника дискуссии использовался факт, что структура информационного текста принципиально отличается от структуры внушающего (манипулирующего) текста и характеризуется отсутствием намеренной ритмизации его лексических и фонетических единиц [1].

Это означает, что некоторые звукосочетания способны не только вызывать определенные эмоции, но и могут восприниматься в качестве определенных образов. Например, в сочетаниях буква «и» с указанием предмета обладает свойством «уменьшения» объекта, перед которым (или в котором) она явно доминантно присутствует. Также, звук «о» производит впечатление мягкости и расслабленности. Преобладание звуков «а» и «э», как правило, ассоциируется с эмоциональным подъемом.

Когда человек, обращающийся в службу поддержки пользователей, четко понимает, что он хочет узнать, например, информацию о способах оплаты, то его сообщение будет носить чисто информационный характер. Если же человек возмущен, например, несанкционированным съемом средств с его банковской карты или навязыванием какой-либо платной услуги, то его обращение будет носить манипулятивный характер («верните деньги!» или «что за дела?!»).

Таким образом, имея достаточно большую статистику обращений, можно по частотному дисбалансу гласных букв в тексте отличить манипулятивное обращение от информационного.

Основное отличие такого подхода от, например, словарной оценки степени эмоциональности текста [2, 3, 12], является его «низкоуровневость». Оцениваются не слова целиком, а частотные характеристики слов. При этом уменьшается вероятность ошибок классификации, поскольку сымитировать частотный дисбаланс гласных в словах намного сложнее, чем подобрать слова из словаря. Это особенно актуально, когда запросы в службу поддержки отправляются не реальными людьми, а программными ботами с целью паралича ее работы.

Для определения частотных характеристик гласных букв в обращениях применяется следующая формула:

$$f_{k,p} = \frac{N_{k,p}}{N_p} \quad (1)$$

где k – символ, для которого делается расчет,

$N_{k,p}$ – кол-во символов k в сообщении пользователя p,

N_p – общее количество символов в запросе пользователя p.

Авторами были выделены буквы, а также знаки, которые наиболее значимы для оценки эмоционального фона (табл. 1).

Таблица 1

Список полей для анализа

Поле	Комментарий	Поле	Комментарий
M_p	M_p-количество сообщений для p-го пользователя	$f_{у,p}$	Частота встречаемости символа «у» для p-го пользователя
L_p	Средняя длина сообщения для p-го пользователя	$f_{э,p}$	Частота встречаемости символа «э» для p-го пользователя
$f_{а,p}$	Частота встречаемости символа «а» для p-го пользователя	$f_{ю,p}$	Частота встречаемости символа «ю» для p-го пользователя
$f_{е,p}$	Частота встречаемости символа «е» для p-го пользователя	$f_{я,p}$	Частота встречаемости символа «я» для p-го пользователя
$f_{и,p}$	Частота встречаемости символа «и» для p-го пользователя	$f_{!,p}$	Частота встречаемости символа «!» для p-го пользователя
$f_{о,p}$	Частота встречаемости символа «о» для p-го пользователя	$f_{?,p}$	Частота встречаемости символа «?» для p-го пользователя

Поскольку мы не можем заранее знать, какие категории обращений могут быть, т.к. степень эмоциональной окраски текста может разная для разных пользователей, то также, как и в работе [6], авторами использовались самоорганизующиеся нейронные сети – сети Кохонена [5, 8].

Данный вид сетей позволяет выявлять кластеры (группы) входных векторов, обладающих некоторыми общими свойствами. В нашем случае это сообщения, имеющие схожую эмоциональную окраску. На рисунке 1 приведен пример группировки текстов по эмоциональному признаку. Здесь приведены примеры разбиения на группы по нескольким параметрам из таблицы 1. Каждый шестиугольник – это проекция конкретного обращения пользователя на карту Кохонена. На рисунке хорошо видно, какие параметры являются более значимыми для разбиения обращений на группы, а какие малозначимыми.

После того, как из сообщения пользователя будет выделена эмоциональная составляющая (в нашем случае это номер кластера на карте Кохонена), необходимо проанализировать смысловую часть сообщения. Для этого определяем список категорий, по которым будет распределяться текст. Это могут быть либо отделы, либо даже отдельные специалисты. Например, категория «юридический отдел» или категория «отдел доставки». Переводим текст образца запроса в вещественное пространство признаков, используя модель bag-of-words (мешок слов) [9, 10, 11]. А для каждой установленной категории создаем ключевые словари. Сначала ставится 0, а если встречается слово или его синоним из словаря, то добавляем 1. Соответственно, чем больше единиц, тем выше шанс принадлежности текста к одной из определенных категорий.

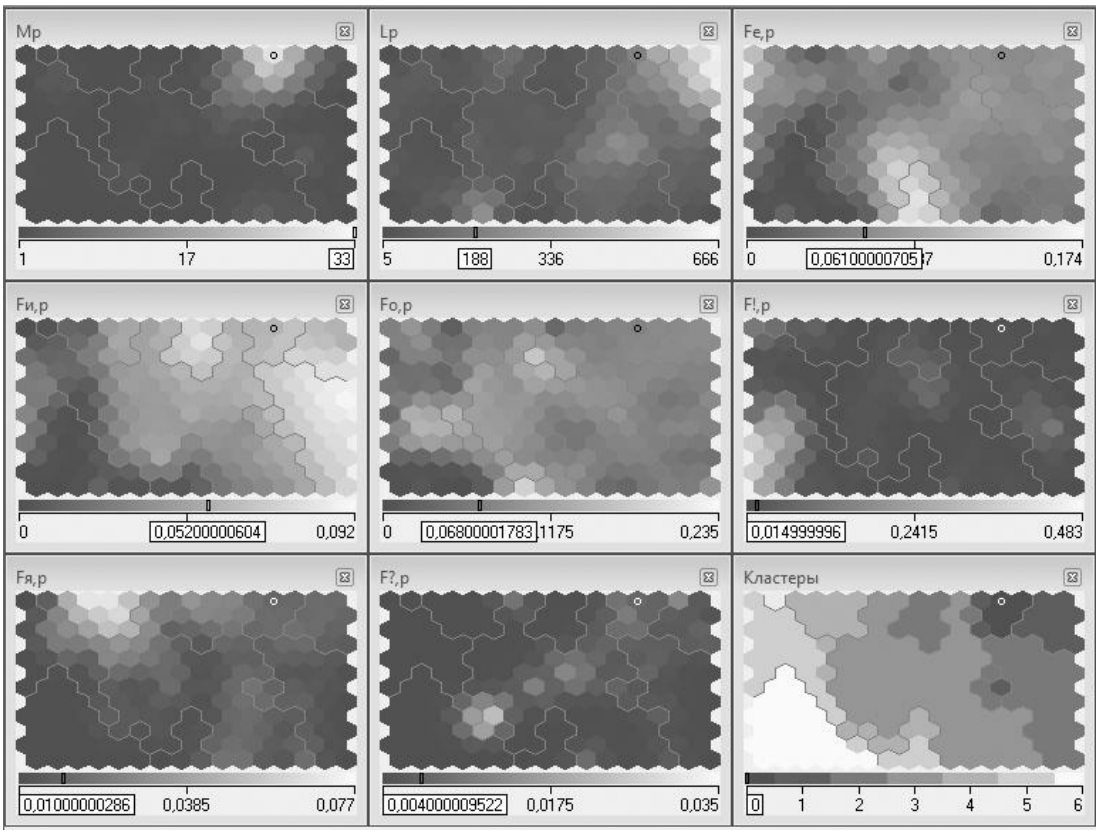

Рисунок 1. Группировка текстов в проекции на карты Кохонена

Для наглядности рассмотрим работу алгоритма на следующих примерах текстовых образцов, поступающих сотруднику технического отдела: «Я вовремя заплатил. Почему на моём компьютере подключен интернет, но он не работает?!».

Составим небольшие словари для нескольких категорий:

«Словарь_1» (технический отдел) - пропускная способность, компьютер, роутер, работоспособность, работать, интернет.

«Словарь_2» (подключение и тарифы) - пакет, тариф, трафик, план, скорость, подключение, абонент, интернет.

Также добавим еще один запрос 2: «Где мой интернет?!».

Составленные вектора принадлежности указаны в таблице 2.

Таблица 2

Значения принадлежности двух образцов

Запрос 1	Словарь_1	Словарь_2	Запрос 2	Словарь_1	Словарь_2
вовремя	0	0	где	0	0
заплатил	1	1	интернет	1	1
интернет	1	1	мой	0	0
компьютере	1	0			
моем	0	0			
на	0	0			
не	0	0			
но	0	0			
он	0	0			
подключен	0	1			
почему	0	0			
работает	1	0			
я	0	0			

В итоге для запроса 1 получаем вектор:

$$\overrightarrow{X_1} = \begin{cases} 0111000000010 \\ 0110000001000 \end{cases} = 0111000001010,$$

и большую вероятность быть распознанным как запрос в технический отдел.

Из запроса 2 видно, что он может быть равновероятно распознан, как и запрос в технический отдел, так и запрос в отдел подключений и тарифов:

$$\overrightarrow{X_2} = \begin{cases} 010 \\ 010 \end{cases}$$

Для итоговой классификации сообщений объединяются вектор принадлежности из таблицы 2 и номер кластера из карты Кохонена (эмоциональная составляющая) в один входной вектор, как показано на этом примере: 01110000010106, где последняя цифра – это номер кластера.

Полученный вектор уже можно использовать как вход для итогового классификатора, в качестве которого авторами предлагается использовать наиболее распространенную разновидность нейронных сетей – многослойные персептроны.

Обобщенная схема работы алгоритма классификации сообщений пользователей представлена на рисунке 2.

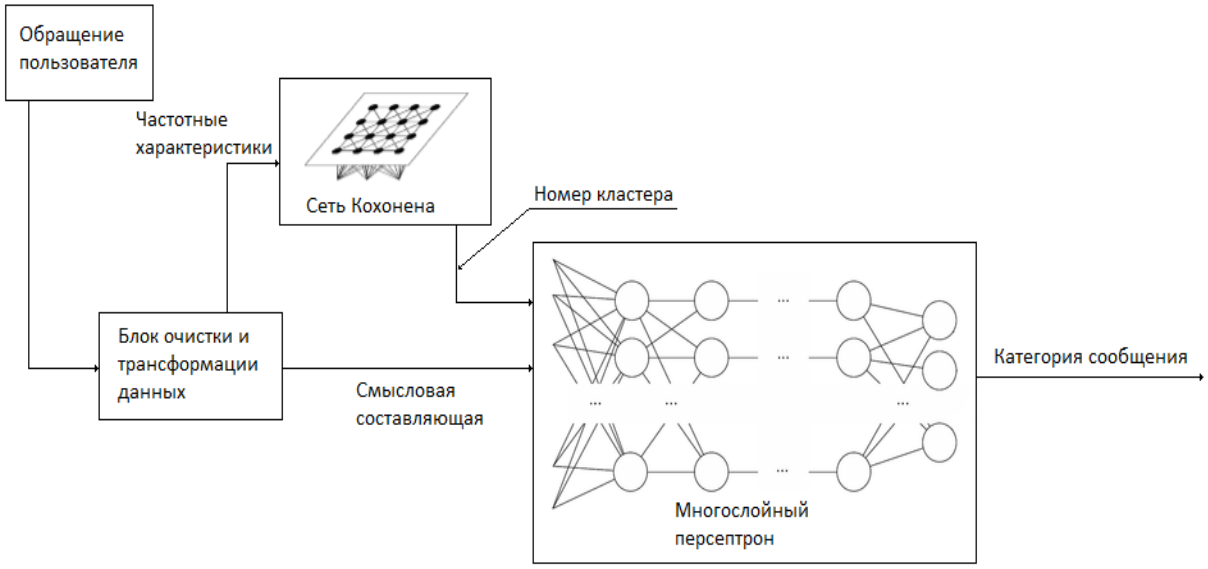

Рисунок 2. Обобщенная схема работы алгоритма

Классификация обращений пользователей будет выглядеть следующим образом:

1. Текст обращения подается на блок очистки и трансформации данных, где в соответствии с формулой 1 и таблицей 1 вычисляются частотные характеристики текста обращения. Кроме того, блок трансформации, используя модель «мешок слов», выдает вектор принадлежности аналогично примеру, приведенному в таблице 2.
2. Частотные характеристики пропускаются через сеть Кохонена с целью определения группы (номер кластера), к которой относится обращение по эмоциональному признаку.
3. Номер кластера объединяется с вектором принадлежности, полученным с помощью модели «мешок слов» (добавляется еще одна компонента вектора), и результирующий вектор подается на вход многослойного персептрона.
4. Выход персептрона определяет категорию сообщения.

Практическая часть. Для корректной работы описанного алгоритма необходимо предварительно произвести обучение нейронных сетей.

Авторами использовалась обучающая выборка, состоящая из 800 обращений пользователей (данные были предоставлены ООО «Лаборатория нейросетевых технологий»), и еще 90 обращений использовались для тестирования результатов.

В качестве инструментария использовался аналитический пакет Deductor Academic.

Сеть Кохонена и персептрон обучаются независимо друг от друга.

При обучении сети Кохонена использовались следующие настройки:

1. Определение количества кластеров автоматическое;
2. Скорость обучения в начале 0,35, а в конце 0,001;

3. Радиус обучения в начале 5, а в конце 0,1;
4. Максимальное количество эпох обучения 1000.

В таблице 3 приведен пример обучающего множества для сети Кохонена.

Таблица 3

Пример обучающего множества

p	M_p	L_p	$f_{а,p}$	$f_{е,p}$	$f_{и,p}$	$f_{о,p}$	$f_{у,p}$	$f_{э,p}$	$f_{!,p}$	$f_{ю,p}$	$f_{я,p}$	$f_{?,p}$
1	1	61	0,092	0,033	0,082	0,066	0,032	0,017	0,000	0,000	0,000	0,000
2	1	20	0,100	0,050	0,050	0,100	0,000	0,000	0,000	0,000	0,000	0,000
3	1	102	0,029	0,078	0,078	0,078	0,010	0,000	0,186	0,000	0,000	0,000
4	1	52	0,096	0,058	0,019	0,068	0,018	0,000	0,058	0,000	0,000	0,000

По итогам обучения было выделено 7 групп (кластеров), т.е. было выявлено 7 уровней эмоционального тона сообщений пользователей.

При обучении персептрона была признана оптимальной следующая структура нейронной сети:

1. Количество слоев – 3
2. Количество нейронов в входном слое – 81
3. Количество нейронов в скрытом слое – 17
4. Количество выходных нейронов – 3

Категория сообщений пользователей кодировалась тремя компонентами, например:

011 – запрос к бухгалтерии;

101 – ошибки в программном обеспечении;

100 – сложности в установке ПО и т.п.

Алгоритм обучения персептрона – Resilient Propagation.

Результаты обучения:

Максимальная ошибка на обучающем множестве – 0,023;

Средняя ошибка на обучающем множестве – 0,0098;

Максимальная ошибка на тестовом множестве – 0,12;

Средняя ошибка на тестовом множестве – 0,05;

Количество эпох обучения – 118.

Учитывая тот факт, что эмоциональная составляющая входного вектора представлена одним компонентом, а смысловая 80 компонентами, в момент начала обучения весовые коэффициенты связей, ведущих к входу с номером кластера (эмоциональная составляющая), принудительно были выставлены на порядок больше, чем весовые коэффициенты связей, ведущих к остальным входам.

Такая процедура была сделана для того, чтобы смысловая и эмоциональная части обращений в момент запуска обучения не имели преимуществ друг перед другом.

Это связано с тем, что состояние нейрона определяется по формуле:

$$S = \sum_i x_i w_i \qquad (2)$$

где x_i – i-й компонент входного вектора, а w_i – это весовой коэффициент связи, ведущей к i-ому входу нейрона.

Поскольку количество компонентов входного вектора в нашем примере было равно 81, а в момент начала обучения веса в сети инициализируются случайными числами вблизи ноля, то компонент эмоциональной составляющей не имеющий принудительно выставленного большого веса, внес бы несущественный вклад в итоговую сумму (во всяком случае, в начале обучения). Это могло привести к увеличению времени обучения.

Заключение. В ходе выполнения работы авторами был разработан и апробирован алгоритм классификации текстовых обращений пользователей интернет-ресурсов на основе анализа смысловой и эмоциональной составляющих текста с использованием нейронных сетей.

Литература

1. Батура, Т.В. Методы автоматической классификации текстов / Т.В. Батура // Международный научно–практический журнал. — 2017. — №1(3). — С. 85 — 99.
2. Клековкина, М. В., Котельников, Е.В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // RCDL-2012, Переславль-Залесский, Россия: конференция. — 2012.
3. Меньшиков, И. Анализ тональности текста на русском языке при помощи графовых моделей // УРФУ, Екатеринбург, Россия: конференция. — 2012.
4. Осовский, С. Нейронные сети для обработки информации [пер. с польского И.Д. Рудинского]. М.: Финансы и статистика, 2002 344с.
5. Солдатова, О.П., Чайка, П.Д. Исследование эффективности решения задачи классификации распределённой гибридной сетью Кохонена // Труды международной научно-технической конференции «Перспективные информационные технологии (ПИТ 2014)»

(Самара, СГАУ, 30 июня–2 июля 2014 г.). Самара: Издательство Самарского научного центра РАН, 2014 с.170-173.
6. Филимонов, А.В. Применение нейронных сетей для выявления троллей в социальных сетях / А.В. Филимонов, А.В. Осипов, А.Б Климов // Нейрокомпьютеры: разработка, применение. — 2015. — № 8. — С. 87— 92.
7. Хайкин, С. Нейронные сети. Полный курс / 2-е изд., испр.: Пер. с англ. – М.: ООО «И. Д. Вильямс», 2006. – 1104 с.
8. Kohonen, T. Self-organizing maps // Springer Science & Business Media, 2001 V. 30
9. Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient Estimation of Word Representations in Vector Space. // In Proceedings of Workshop at International Conference on Learning Representations (ICLP) – 2013, [Электронный ресурс] – Режим доступа: http://arxiv.org/abs/1301.3781
10. Mikolov, T., Le, Q. Distributed Representations of Sentences and Documents. // In Proceedings of Workshop at The 31st International Conference on Machine Learning (ICML) – 2014, [Электронный ресурс] – Режим доступа: http://jmlr.org/proceedings/papers/v32/le14.pdf
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. Distributed Representations of Word and Phrases and their Compositionaly. //In Proceedings of Workshop at The Twenty-seventh Annual Conference on Neural Information Processing Systems (NIPS) – 2013, [Электронный ресурс] – Режим доступа: http://arxiv.org/abs/1310.4546
12. Su, F., Markert, K. From Words to Senses: a Case Study in Subjectivity Recognition // Proceedings of Coling, Manchester, UK. — 2008.

УДК 681.5

ПРОЕКТИРОВАНИЕ ЕДИНОГО ИНФОРМАЦИОННОГО ПРОСТРАНСТВА ТЕЛЕКОММУНИКАЦИОННОЙ КОМПАНИИ

Власова Эльвира Андреевна (elvirus97@mail.ru)
Сизова Ольга Владимировна
ФГБОУ ВО «Ивановский государственный химико-технологический университет»

В работе рассматривается аспект автоматизации деятельности телекоммуникационной компании. В ходе осуществления проекта создается единое информационное пространство организации с целью повышения эффективности ее деятельности. В работе был проведен анализ ИТ-архитектуры рассматриваемой организации, который позволил выявить главный недостаток существующего информационного пространства организации. Сотрудники не имеют полного доступа к информации, которая им необходима для осуществления их должностных обязанностей. Для разрешения возникшей ситуации в работе была построена модель единого информационного пространства в виде web-сервиса и разработана структура прикладного решения.

Ключевые слова: информационные ресурсы, информация, информационное пространство, доступ к информации, модель, структура, интеграция.

Целью любой организации является эффективное использование информационных ресурсов и их грамотная реализация. В настоящее время индустрия информационных технологий выведена в качестве одного из ведущих стратегических направлений.

В большинстве организаций одним из главных, используемых для деятельности инструментов, являются информационные системы (ИС). А в крупных компаниях, такой, как МТС, например, систем очень много. Если все системы интегрированы в единую информационную среду, то это только повышает эффективность деятельности, если же нет – то приводит к снижению производительности труда, доставляет неудобство сотрудникам и замедляет процесс решения ежедневных задач.

Объединение информационных ресурсов на основе взаимодействия информационных систем выведет компанию на уровень корпоративных информационных ресурсов. Такое объединение называется создание единого информационного пространства организации [1].

Концепция единого информационного пространства реализует идею полной комплексной автоматизации управления организацией.

Единое информационное пространство (ЕИП) представляет собой совокупность баз и банков данных, технологий их ведения и использования, информационно-телекоммуникационных систем и сетей, функционирующих на основе единых принципов и по общим правилам, обеспечивающим информационное взаимодействие организаций и граждан, а также удовлетворение их информационных потребностей [2].

При использовании в ЕИП прикладных программ, в каждой из информационных систем часть методов обработки данных реализуется в виде приложений, доступных из других информационных систем. Например, при взаимодействии двух ИС первая пользуется сервисами,