УДК 519. 253, 378.147.8
DOI: 10.12737/22128

С.М. Иванова, З.В. Ильиченкова

СИСТЕМА ПОИСКА И АНАЛИЗА ДОСТОВЕРНОЙ ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ

Рассмотрены вопросы, связанные с необходимостью для пользователей нахождения различной информации во всемирной сети Интернет и возможным определением достоверности этой информации. Описана технология поиска в сети, показаны поисковые инструменты и определены правила эффективного поиска. Разработан критерий достоверности сайта на основе методов нечёткой логики и показан способ оценки интегральной достоверности сайта.

Ключевые слова: поиск в сети Интернет, достоверность информации, поисковые инструменты, нечёткая логика, интегральная достоверность сайта.

S.M. Ivanova, Z.V. Ilichenkova

SYSTEM OF RETRIEVAL AND ANALYSIS OF RELIABLE INFORMATION IN INTERNET NETWORK

Now in the world a problem of reliable information retrieval in the Internet network is especially critical. It is defined by increased data capacity and by the absence of control over data placement. It is offered to determine information reliability by methods of fuzzy logic at the correct formation of retrieval request. For the estimate of completeness, truth and integrity of data obtained it is offered to rely upon the analysis of information from other pages of a site found. In accordance with this the notions of topical closeness of site pages and reliability of information presented in the remaining sections. To organize an output are compiled the rules for the integral reliability output of a site page. For example, if information on a page does not coincide with analyzed one and is reliable then one can suppose that related information is reliable. Further, with the use of Mamdani controller is carried out a defuzzification. A technology for a retrieval request organization is offered to increase an information content retrieval. The technologies offered for retrieval and analysis of information allow increasing effectiveness and a retrieval rate of reliable, integrated and true information in the Internet network.

Key words: retrieval in the Internet network, information reliability, searching tools, fuzzy logic, site integral reliability.

С каждым годом объемы различной информации, находящейся во всемирной системе объединенных компьютерных сетей Интернет, увеличиваются в геометрической прогрессии, поэтому вероятность нахождения необходимой информации в Интернете близка к единице. Однако возникает проблема нахождения необходимой достоверной информации. Всемирная паутина объединяет миллионы компьютеров, множество разных сетей, число пользователей увеличивается на 30-50% ежегодно. Безусловно, Интернет увеличивает возможности получения дополнительных знаний, помогает грамотно организовывать самообразовательную деятельность. Но, к сожалению, исследования социологов показывают, что среднестатистический пользователь готов потратить на поиск не более 15 минут своего времени. Поэтому очень важно, с одной стороны, организовать поиск информации таким образом, чтобы используемая поисковая система выдавала информацию, наиболее близкую к запрашиваемой, а с другой стороны, уметь оценивать достоверность тех данных, которые были найдены. Таким образом, для грамотной организации поиска достоверной информации в сети Интернет необходим комплексный подход к проблеме [1; 2; 6].

Данная проблема является важной не только в самообразовательной деятельности. Особую важность она имеет при организации самостоятельной работы обучающихся, в том числе и при электронном

обучении [3]. Согласно государственной программе РФ «Развитие образования» на 2013-2020 годы [5] и федеральным государственным образовательным стандартам (ФГОС), в данный момент большое внимание уделяется самостоятельной работе в процессе обучения студентов. Не последнюю роль в этой работе занимает получение дополнительной информации из Интернета.

Определение достоверности информации

Достоверность информации определяется:

- полнотой;
- целостностью;
- истинностью.

Предлагается рассмотреть следующий пример. При создании модели устройства, которое может находиться в конечном количестве состояний и переходить из одного состояния в другое в фиксированные моменты времени, требуется решить систему линейных алгебраических уравнений (СЛАУ). При этом характеристика данных такова, что не является очевидным, какой из методов решения будет оптимальным. То есть необходимо найти информацию о наиболее адекватном задаче методе решения СЛАУ с учётом задаваемых ограничений.

Нужно отметить, что большинство сайтов предлагают два самых популярных и широко распространенных метода: метод Гаусса и метод Крамера. Если рассмотреть первые пятьдесят ссылок в поисковой системе Яндекс, то в половине встречаются различные модификации метода Гаусса, в семи ссылках есть упоминание о методе Крамера, в трех встречается метод итераций и его модификации, один раз упоминался метод Холецкого, в остальных предлагалось просто решить СЛАУ on-line без упоминания способа ее решения. Другие способы решения СЛАУ вообще не упоминались. Таким образом, в данном случае, при использовании Интернета как дистанционного образовательного ресурса, студенту придется потрудиться, чтобы найти полную информацию по вопросу решения СЛАУ. Кроме этого все приведенные выше методы имеют определенные ограничения применимости, а это тоже достаточно редко указывается в сопроводительных статьях. То есть зачастую информация, получаемая пользователем по запросу, не удовлетворяет критерию полноты.

Целостность информации определяется тем, насколько сильно она была искажена при ее передаче, приеме и хранении. Иногда при перепечатке каких-либо фактов с одного сайта на другой допускаются ошибки. Про орфографические ошибки и опечатки в Интернете не писал только ленивый, но всё чаще встречаются ошибки и фактического характера [8]. Отчасти это связано с тем, что помещать свои статьи во Всемирной паутине может любой пользователь, не заботясь о качестве и достоверности изложенной информации. Хотя придется отметить, что этим все чаще страдают теперь практически любые источники информации. Чтобы избежать подобных оплошностей, приходится сравнивать информацию с нескольких сайтов.

Определение истинности и точности информации также является очень сложной задачей для пользователя Интернета. Чтобы обезопасить себя от ложной информации, следует использовать комплексные методы. Например, при получении информации следует посмотреть на адрес сайта или портала. В Интернете есть зарезервированные части адресов. Например, если адрес оканчивается на -gov.ru – это ресурс правительственной организации; -ac.ru – это ресурс академической организации (НИИ или вуза); -edu.ru – это ресурс образовательных организаций. Такие адреса выдаются только специальным организациям, и вероятность ошибки на таком сайте ничтожно мала. Еще один способ – проверить, есть ли ссылки на авторов статей, указаны ли источники информации, являются ли данные источники авторитетными.

Для определения достоверности информации рассмотрим лишь один из критериев – её истинность. Каким образом

можно проверить самостоятельно истинность найденной информации, если исключить возможности авторитетности адреса сайта и консультации у специалиста. Предлагается проверить информацию, используя свои базовые знания по этому предмету. Продолжим рассматривать означенный выше пример решения СЛАУ. Выбираем сайт, где прописываются несколько способов решения, в том числе и необходимый. Если пользователь разбирается хотя бы в одном методе, то можно проверить истинность известного метода. Тогда если в изложении нет ошибок, то возможно использовать этот сайт для изучения других методов, хотя такой способ тоже не дает стопроцентной гарантии. Можно использовать и другой вариант: взять тестовый пример с заранее известным ответом и проверить на этом примере описанные методы решения.

Безусловно, что для оценки критерия достоверности сайта целесообразно использовать основные понятия нечеткой логики [4]. Для этого определяем субъективно понятия:

- близкие результаты (это в том случае, если результаты тестирования совпали);
- похожие результаты (если результаты отличаются на малую величину ввиду погрешности округления);
- есть общее в результатах (если погрешности достаточно велики);
- нет совпадений.

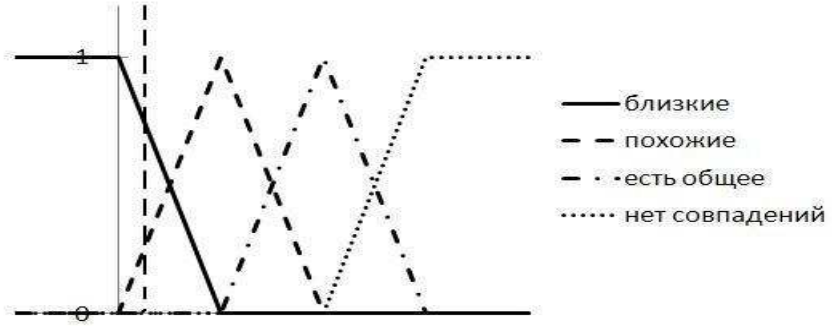

Рис.1. Характеристическая функция полученных результатов для определения близости метода решения

События «близкие», «похожие», «есть общее» и «нет совпадений» (рис. 1) образуют полную группу. По оси OY расположена шкала истинности (изменяется от 0 до 1). Вертикальный пунктир показывает, что в данном случае проверка работы различных методов дает результат между близким и похожим. Тогда на основе полученных данных можно сделать вывод о достоверности сайта и возможности дальнейшего его использования.

На рис. 2 события «недостоверный», «что-то похожее есть», «можно учесть», «достоверный» образуют полную группу. По оси OY расположена шкала истинности (изменяется от 0 до 1). Вертикальный пунктир показывает, что данный сайт имеет высокую степень достоверности.

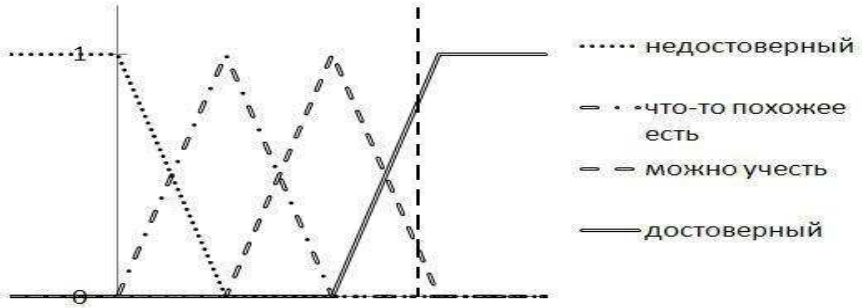

Рис. 2. Характеристическая функция достоверности страницы сайта

Для принятия решения о достоверности предлагается определить правила нечёткого вывода определения интегральной достоверности страницы сайта относительно искомой информации следующим образом. Если информация на странице близка к изучаемой и является достоверной, то можно предполагать, что искомая информация классифицируется как полностью достоверная. Если информация на странице не совпадает с изучаемой и является достоверной, то искомая информация, возможно, является достоверной. Полностью все правила вывода представлены в таблице.

Таблица

Правила вывода интегральной достоверности страницы сайта

условная близость→ достоверность страницы ↓	близкие	похожие	есть общее	нет совпадений
недостоверный	возможная	сомнительная	условная	недостоверная
что-то похожее	вероятная	возможная	сомнительная	условная
можно учесть	определённая	вероятная	возможная	сомнительная
достоверный	полная	определённая	вероятная	возможная

С помощью контроллера Мамдани [7] проводится дефаззификация. Например, при исходных данных условной близости (рис. 3) и определённой достоверности (рис. 4) результат определяется как центр тяжести нечёткого множества (обозначен белой точкой на рис. 5).

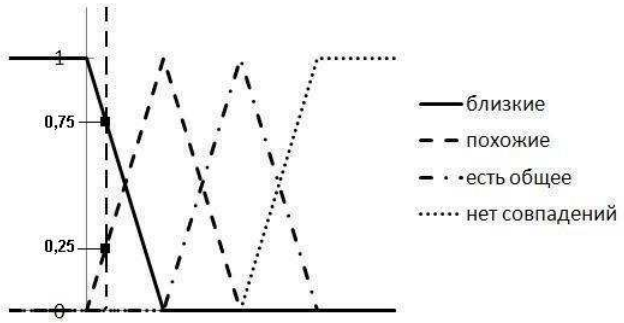

Рис. 3. Определённое значение близости

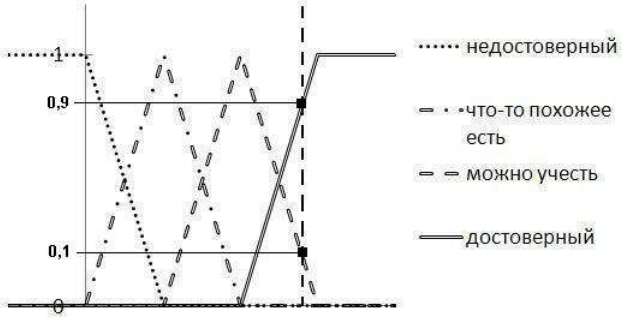

Рис. 4. Определённое значение достоверности

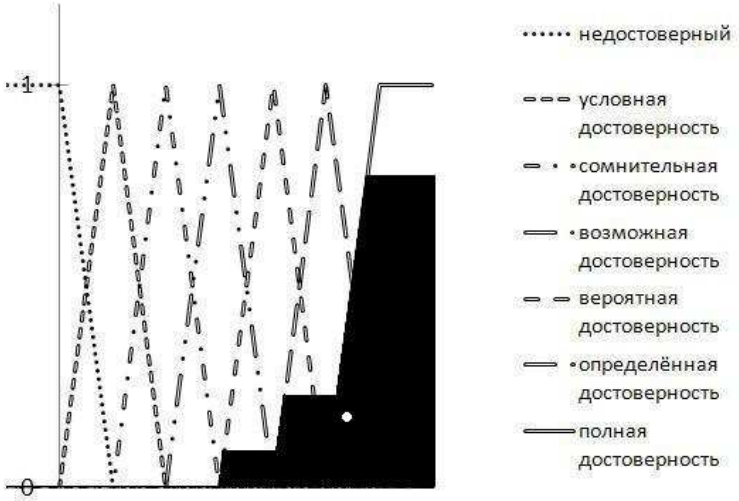

Рис. 5. Нечёткое множество одной страницы и его дефаззификация

Таким образом, решением контроллера нечеткой логики является нечеткое множество (для определения интегральной достоверности страницы сайта). Интегральную достоверность сайта предлагает-

ся определять как среднеквадратичную из всех отдельных достоверностей.

Следовательно для определения достоверности информации, изложенной на сайте, необходимо иметь:

- базовые знания по данному предмету;
- возможность посоветоваться с профессионалом по данному вопросу;
- возможность проверить информацию с помощью других авторитетных ресурсов.

Хотя надо отметить, что, соблюдая эти правила, все равно нельзя застраховать себя от ложной информации на сто процентов, можно лишь уменьшить вероятность ее получения.

Технологии поиска информации в Сети для повышения её достоверности

Предложенный выше метод определения достоверности информации хорошо работает в случае, когда пользователь может довольно быстро определить необходимые первичные характеристики полученной информации. При этом важную роль играет такой фактор, как близость полученной информации к искомой.

Для поиска необходимой информации, во-первых, нужно найти её адрес. Для этого существуют специализированные поисковые серверы (роботы индексов (поисковые системы), тематические интернет-каталоги и др.). Поисковые системы обычно состоят из трех компонентов: агента, который собирает информацию в сети Интернет; базы данных, содержащей всю имеющуюся информацию; поискового механизма, который люди используют как интерфейс для взаимодействия с базой данных.

Средства поиска и структурирования дают возможность найти необходимую информацию. Поисковыми инструментами называют программное обеспечение, которое позволяет обеспечить наиболее оптимальный и качественный поиск информации для пользователей Интернета. Поисковые инструменты размещаются на специальных веб-серверах, каждый из которых выполняет определенную функцию: анализ веб-страниц; поиск информации по запросу пользователя; обеспечение удобного интерфейса для поиска информации и просмотра результата поиска. Способы работы с различными поисковыми инструментами, практически не отличаются.

Важным является тот факт, что так как порядок представления ссылок в результате поискового запроса определяется исходя из семантического анализа, проводимого поисковой системой, то точность ответа во многом определяется совпадением логики запроса и организации поиска.

Организация поиска информации в сети Интернет устроена следующим образом: пользователь набирает ключевую фразу и активизирует поиск, тем самым получает подборку ссылок на различные сайты по своему запросу. Данный список документов формируется так, чтобы в начале списка стояли те ссылки, которые наиболее соответствуют запросу пользователя. Каждый из поисковых инструментов использует различные критерии ранжирования документов при анализе результатов поиска. Таким образом, на одинаковый запрос можно получить различные результаты поиска в разных браузерах. Как правило, самый терпеливый пользователь просматривает первые две-три страницы, а это порядка тридцати ссылок. По результатам поиска для пользователя имеет большое значение, насколько необходимыми и полезными для него являются эти ссылки. Пользователю предлагаются два способа поиска: простой поиск и расширенный поиск (с использованием специальной формы запроса и без нее). Простой запрос отражает большое количество ссылок, и их просмотр занимает достаточно много времени, так как в список попадают ссылки, содержащие хотя бы одно слово, введенное при запросе.

Как показывает статистика, только двадцать процентов пользователей могут грамотно сформулировать запрос. При перечислении через запятую ключевых слов в поисковике можно получить лишь общую информацию об интересующем вас предмете. Вспомнив все тот же пример о различных способах решения СЛАУ, пользователю в списке ссылок сразу следует

отмести те, где встречающиеся словосочетания связаны с получением конкретного расчета или on-line. Полезно в этом случае выбирать ссылки, содержащие слово «пример», так как очень часто одна теория, освещающая данный вопрос, не позволяет полностью разобраться в необходимой информации. Если у студента имеются базовые знания по этому вопросу, то ему гораздо проще разобраться в потоке информации. Знания в предметной области позволяют заметно сократить время поиска: при просмотре сайта ему достаточно всего 10-15 секунд.

Большое значение имеет правильность формулировки запроса: ее точность, конкретность, соответствие правильной терминологии, правильность сочетаний при использовании различных частей речи, более общая или более конкретная формулировка запроса. При получении полного набора ссылок необходимо выбрать ту, которая позволит наиболее полно и точно ответить на заданный вопрос.

К сожалению, все чаще встречаются сайты с достоверной информацией, но изложенной сумбурно. В таких сайтах сложно разобраться, и, как правило, они не пользуются популярностью у читателей. Структурированная информация всегда легче воспринимается, запоминается и анализируется.

Таким образом, можно сформулировать несколько правил поиска информации в Интернете:

1. Необходимо определиться с тем, какую информацию вы собираетесь найти, т.е. четко сформулировать тему запроса. Это позволит вам получить истинную информацию.

2. Необходимо верно написать ключевые слова, небуквенные символы. Каждая поисковая система имеет свою форму составления запросов. Несмотря на то что принцип один, могут отличаться используемые операторы.

3. Если информацию не удается найти в привычном поисковике - на Google (первая в мире по популярности система, обрабатывает 41345 млн запросов в месяц) или Яndex (четвёртая среди поисковых систем мира по количеству обработанных поисковых запросов (4,84 млрд) очень популярная в России), попробуйте другие поисковые системы. Кроме того, можно воспользоваться услугами расширенного поиска. Поиск в различных поисковых системах позволяет увеличить полноту и целостность информации.

4. Можно использовать знаки «+» или «-» для повышения эффективности поиска, тем самым подчеркивая, что необходимо добавить, а что исключить из запроса.

5. Важно иметь некоторые базовые знания по данному вопросу, чтобы можно было найти сайт с достоверной информацией (хотя бы при беглом просмотре).

Предлагаемые технологии поиска и анализа информации позволяют повысить скорость и эффективность поиска достоверной - полной, целостной и истинной - информации в сети Интернет. Так как поисковые системы не производят самостоятельно информацию, а являются лишь посредником между пользователем и сайтом, то оценивать достоверность полученных данных следует пользователю.

СПИСОК ЛИТЕРАТУРЫ

1. Васильева, А.К. Разработка рекомендаций по внедрению практико-ориентированного обучения по программам Академического портфеля «Вуз-Организация» / А.К. Васильева // Вестник МГТУ «Станкин». - 2015. - № 2 (33). – С. 98-102.
2. Иванова, С.М. Инновационный подход к восстановлению и фильтрации сигналов в линейных динамических системах / С.М. Иванова // Вестник МГТУ «Станкин». – 2009. - № 3. - С. 83-87.
3. Погорелов, Д.Ю. Моделирование динамики систем тел с использованием ПК «Универсальный механизм»: текущее состояние и перспективы развития / Д.Ю. Погорелов // XI Всероссийский съезд по фундаментальным проблемам теоретической и прикладной механики: сб. докл. - 2015. – С. 3027-3029.

4. Рекомендации по оцениванию надёжности интернет-ресурсов. - URL: http://www.pandia.ru/text/78/227/92345.php (дата обращения: 08.04.16).
5. Распоряжение правительства Российской Федерации № 792-р от 15 мая 2013 года. - URL: http://xn--80abucjiibhv9a.xn--1ai/%D0%B4%D0%BE%D0%BA%D1%83%D0%BC%D0%B5%D0%BD%D1%82%D1%8B/3409/%D1%84%D0%B0%D0%B9%D0%BB/2228/13.05.15-D0%93%D0%BE%D1%81%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B0-D0%A0%D0%B0%D0%B7%D0%B2%D0%B8%D1%82%D0%B8%D0%B5_%D0%BE%D0%B1%D1%80%D0%B0%D0%B7%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F_2013-2020.pdf (дата обращения: 08.04.16).
6. Саркисова, И.О. «Перевернутое обучение» как метод оптимизации процесса подготовки IT-специалистов / И.О. Саркисова // Перспективы развития информационных технологий. – 2015. - № 27. – С. 114-121.
7. Харин, А.А. Инновационная инфраструктура как основной элемент национальной инновационной системы / А.А. Харин // Вестник МГТУ «Станкин». - 2015. – С. 12-18.
8. Navara, M. Enhancement of Mamdani Fuzzy Controller. - URL: https://docviewer.yandex.ru/?url=http%3A%2F%2Fwww.eusflat.org%2Fproceedings%2FEUSFLAT_2001%2Fpapers%2F168_Navara.pdf&name=168_Navara.pdf&lang=en&c=57075a3dc0ab (дата обращения: 08.04.16).

1. Vasilieva, A.K. Development of recommendations on introduction of practice-oriented training on curriculum of Academic portfolio “College-Organization” / A.K. Vasilieva // Bulletin of MSTU “Stankin”. - 2015. - № 2 (33). – pp. 98-102.
2. Ivanova, S.M. Innovation approach to restoration and filtration of signals in linear dynamic systems / S.M. Иванова // Bulletin of MSTU “Stankin”. – 2009. - № 3. - pp. 83-87.
3. Pogorelov, D.Yu. Simulation of body system dynamics using “Universal Mechanism” PC: current state and prospects in development / D.Yu. Pogorelov // Proc. of the XI-th All-Russian Congress on Fundamental and Applied Problems of Theoretical and Applied Mechanics. – 2015. – pp. 3027-3029.
4. Recommendations on Estimate of Internet resources reliability. - URL: http://www.pandia.ru/text/78/227/92345.php (address date: 08.04.16).
5. Government Order of the Russian Federation № 792-p dated May 15, 2013. - URL: http://xn--80abucjiibhv9a.xn--1ai/%D0%B4%D0%BE%D0%BA%D1%83%D0%BC%D0%B5%D0%BD%D1%82%D1%8B/3409/%D1%84%D0%B0%D0%B9%D0%BB/2228/13.05.15-D0%93%D0%BE%D1%81%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B0-D0%A0%D0%B0%D0%B7%D0%B2%D0%B8%D1%82%D0%B8%D0%B5_%D0%BE%D0%B1%D1%80%D0%B0%D0%B7%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F_2013-2020.pdf (address date: 08.04.16).
6. Sarkisova, I.O. “Inverse training as method for optimization of IT specialists training / I.O. Sarkisova // *Prospects in Information Technologies Development.* – 2015. – № 27. – pp. 114-121.
7. Kharin, A.A. Innovation infrastructure as basic element in national innovation system / A.A. Kharin // Bulletin of MSTU “Stankin”. - 2015. – pp. 12-18.
8. Navara, M. Enhancement of Mamdani Fuzzy Controller. - URL: https://docviewer.yandex.ru/?url=http%3A%2F%2Fwww.eusflat.org%2Fproceedings%2FEUSFLAT_2001%2Fpapers%2F168_Navara.pdf&name=168_Navara.pdf&lang=en&c=57075a3dc0ab (дата обращения: 08.04.16).

Статья поступила в редколлегию 20.07.2016.
Рецензент: д.т.н., профессор Брянского государственного технического университета Киричек А.В.

Сведения об авторах:

Иванова Светлана Михайловна, к.т.н., доцент кафедры «Информационные системы» МГТУ «Станкин», тел.: 8(910) 437-49-17, e-mail: sm-ivanova@yandex.ru.
Ильиченкова Зоя Викторовна, к.т.н., доцент кафедры «Информационные системЫ» МГТУ «Станкин», тел.: 8(916) 638-34-12, e-mail: zilyichenkova@yandex.ru.

Ivanova Svetlana Mikhailovna, Can.Eng., Assistant Prof. of the Dep. “Information Systems”, MSTU “Stankin”, Phone: 8(910) 437-49-17, e-mail: sm-ivanova@yandex.ru.
Ilichenkova Zoya Viktorovna, Can.Eng., Assistant Prof. of the Dep. “Information Systems”, MSTU “Stankin”, Phone: 8(916) 638-34-12, e-mail: zilyichenkova@yandex.ru.