Мальшаков Г.В., Мальшаков В.Д.

МЕТОДИКА НОРМАЛИЗАЦИИ АЛФАВИТА ПОИСКА ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА ИДЕНТИФИКАЦИИ СУЩНОСТЕЙ ПО ЧАСТОТНЫМ ХАРАКТЕРИСТИКАМ ИХ ДАННЫХ

Аннотация: *Используя частотные распределения данных как их «идентификатор», возможно обнаружить данные одной системы в других предназначенных для взаимодействия системах, тем самым согласовав их работу. В этом случае идентификация сущностей предметной области происходит с помощью алфавита поиска - набора лексем с частотами их использования в данных сущности, располагающихся в записях реляционной базы данных прикладного программного обеспечения.Предметом исследования является методика нормализации алфавита поиска для повышения качества идентификации сущностей предметной области по частотным характеристикам их данных, заключающийся в удалении лексем алфавита входящих в другие лексемы алфавита с аналогичной частотой повтора в данных сущности. В качестве методов исследования использовались системный анализ, теория информации, теория алгоритмов, алгебра логики, теория множеств, сравнительный анализ, методы интеллектуального анализа данных и методы разработки программного обеспечения и баз данных. Экспериментально (на примере 178 сущностей) доказано, что данная методика позволяет в среднем в 5 раз уменьшить объём алфавита поиска, что значительно повышает быстродействие идентификации сущностей по частотным характеристикам их данных.Благодаря уменьшению количества более коротких лексем методика нормализации позволяет уменьшить ошибку распознавания, как показали эксперименты в среднем на 0.02036 на каждую идентификацию.*

Ключевые слова: *нормализация, алфавит, поиск, сущность, частотный анализ данных, корреляция, база данных, программное обеспечение, идентификация, методика*

Abstract: *Using frequency distributions of data as identifier it is possible to find data of one system in other systems intended for interaction and coordinate their work. In this case entity identification of a subject domain is done using the alphabet of search. An alphabet of search is a set of lexemes*

DOI: 10.7256/2305-6061.2015.4.17813

with frequencies of their use in the data, stored as records of a relational database. Object of the research is a technique of normalization of the alphabet of search for improvement of quality of entity identification in a subject domain using frequency characteristics of their data. The technique requires deleting lexemes of the alphabet found in other lexemes of the alphabet with similar frequency of repetition in entity. The methods of the research include the system analysis, the theory of the information, the theory of algorithms, algebra of logic, the theory of sets, the comparative analysis, methods of the intellectual analysis of data and methods of development of the software and databases. The authors prove experimentally (on an example 178 entity), that the given technique allows to reduce the volume of the alphabet of search in 5 times on average, that considerably increases speed of identification entity under frequency characteristics of their data. By reducing the quantity of shorter lexemes the technique of normalization allows to reduce an error of recognition on average by 0.02036 per identification as shown by experiments.
Keywords: *correlation, frequency analysis of data, entity, search, the alphabet, normalization, database, software, identification, method*

Используя частотные распределения данных как их «идентификатор», возможно обнаружить данные одной системы в других предназначенных для взаимодействия системах, тем самым согласовав их работу [1, 2].

Идентификация сущностей по частотным характеристикам их данных необходима для повышения интероперабельности программного обеспечения (ПО) - способности к взаимодействию [3]. Чтобы системы могли между собой взаимодействовать, обмениваясь данными. Это повышает качество жизни конечных пользователей, работающих с этими системами [4].

Данные в системах соответствуют конкретным понятиям прикладной области ПО - сущностям. Каждая сущность имеет набор характеристик - полей сущности. Каждая сущность в жизни воплощается в наборе значений - объектах сущности.

В реляционных базах данных ПО [5, 6] как правило каждая сущность хранится в отдельной таблице. Поля сущности хранятся в отдельных столбцах таблицы сущности. Объект сущности - это сделанная запись в соответствующей таблице сущности.

При идентификации сущностей выполняется расчёт частот встречи лексем поискового алфавита в данных с последующим расчётом коэффициента Пирсона между рассчитанными частотами данных и частотами алфавита лексем поисковой сущности. По величине коэффициента Пирсона (для положительного решения он должен быть > 0.7) принимается решение о соответствии анализируемых данных поисковому алфавиту сущности.

При идентификации огромную роль играет алфавит поиска от него зависит качество (количество ошибок) и скорость распознавания. Алфавит поиска сущности - это набор лексем с частотами их использования в данных объектов сущности. В качестве данных

DOI: 10.7256/2305-6061.2015.4.17813

объекта в исследованиях использовалось ключевое поле имеющее строковый тип (в базе данных поле «Наименование»).

При создании алфавита каждое ключевое слово объекта сущности разбивается на возможные комбинации символов с длинной от 1 до 10 со смещением относительно начала. Если комбинация более одного раза встречается в объектах сущности и её ещё нет в алфавите, то она добавляется в его алфавит. Для получения нормированной частоты встречи лексем в среднем на объект сущности количество повторов делится на количество объектов сущности.

Частотное распределение лексем алфавита для поля «Наименование» биполярного транзистора с изолированным затвором (БТИЗ) представлено на рис. 1.

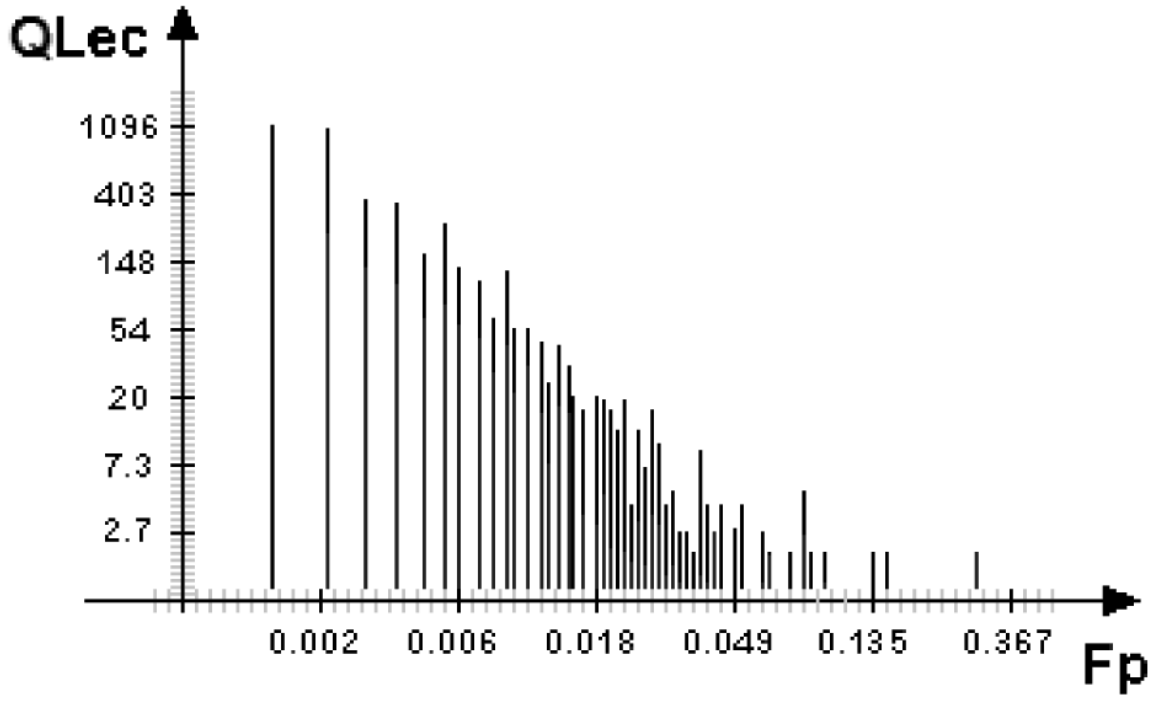

Рис. 1. Частотное распределение лексем алфавита для поля «Наименование» БТИЗ.

QLec - количество лексем, **Fp** - нормированная частота встречи лексем ([количество повторений лексемы] / [количество объектов сущности]) в объектах сущности.

С учётом процедуры получения лексем алфавита существует вероятность вхождения одной лексемы в другую более длинную. При этом если количество их встреч в объектах сущности одинаковое, то одна из этих двух лексем с точки зрения корреляционного анализа бесполезна в силу того, что она дублирует другую, повторяя её частотные характеристики. Для устранения избыточности алфавита введена процедура его нормализации.

При нормализации из алфавита удаляются лексемы входящие в другие лексемы с аналогичной частотой повторов в объектах сущности (рис. 2). При этом удаляются более короткие лексемы для повышения устойчивости к ошибкам ложного срабатывания

DOI: 10.7256/2305-6061.2015.4.17813

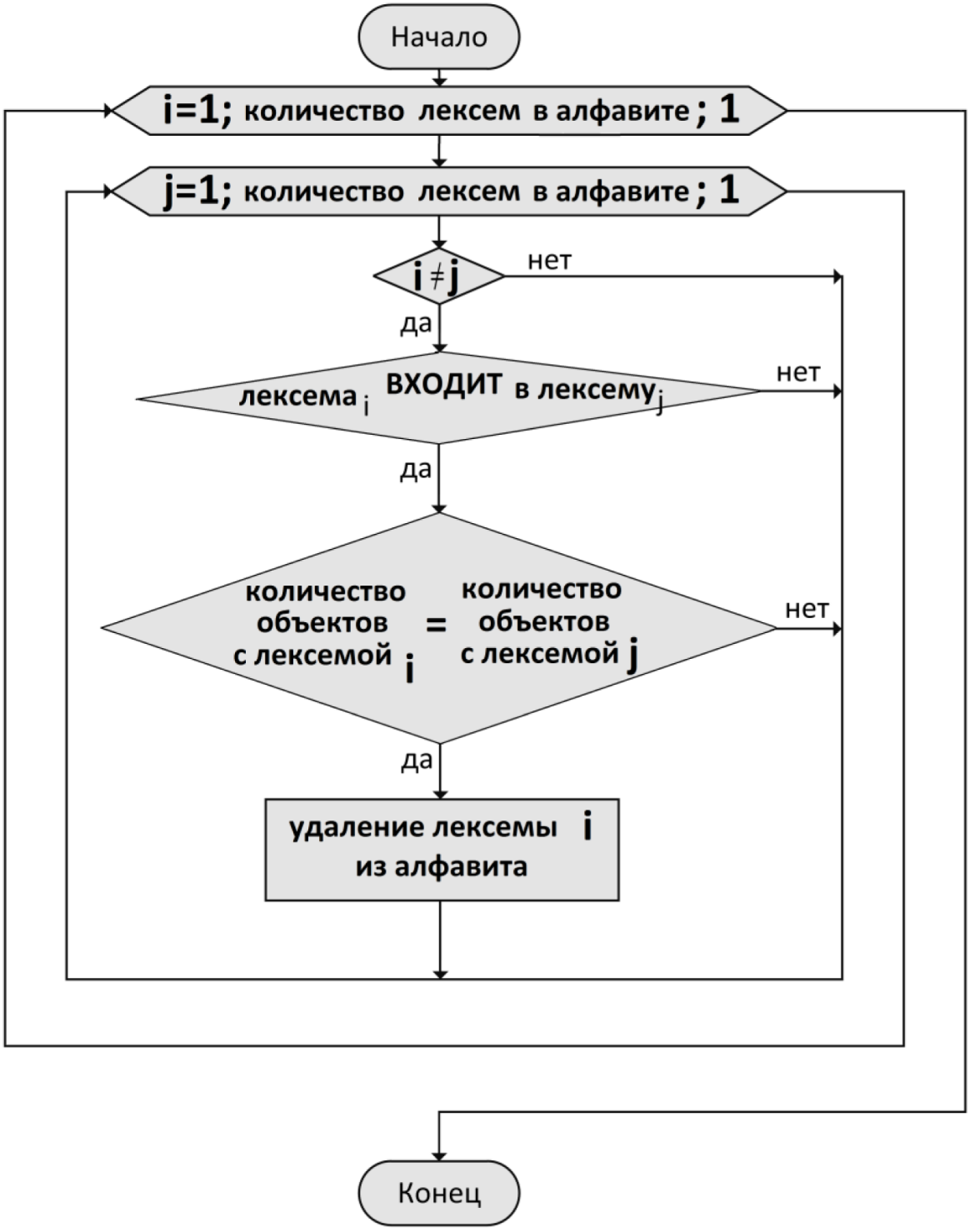

Рис. 2. Алгоритм удаления лексем при нормализации алфавита поиска

DOI: 10.7256/2305-6061.2015.4.17813
При цитировании этой статьи сноска на DOI обязательна

После нормализации в алфавите поиска для поля «Наименование» БТИЗ произошло сокращение числа лексем с 8571 до 2255 (в 3.8 раза). Пять гармоник из 68 спектра частотного распределения лексем сжались до уровня в одну лексему.

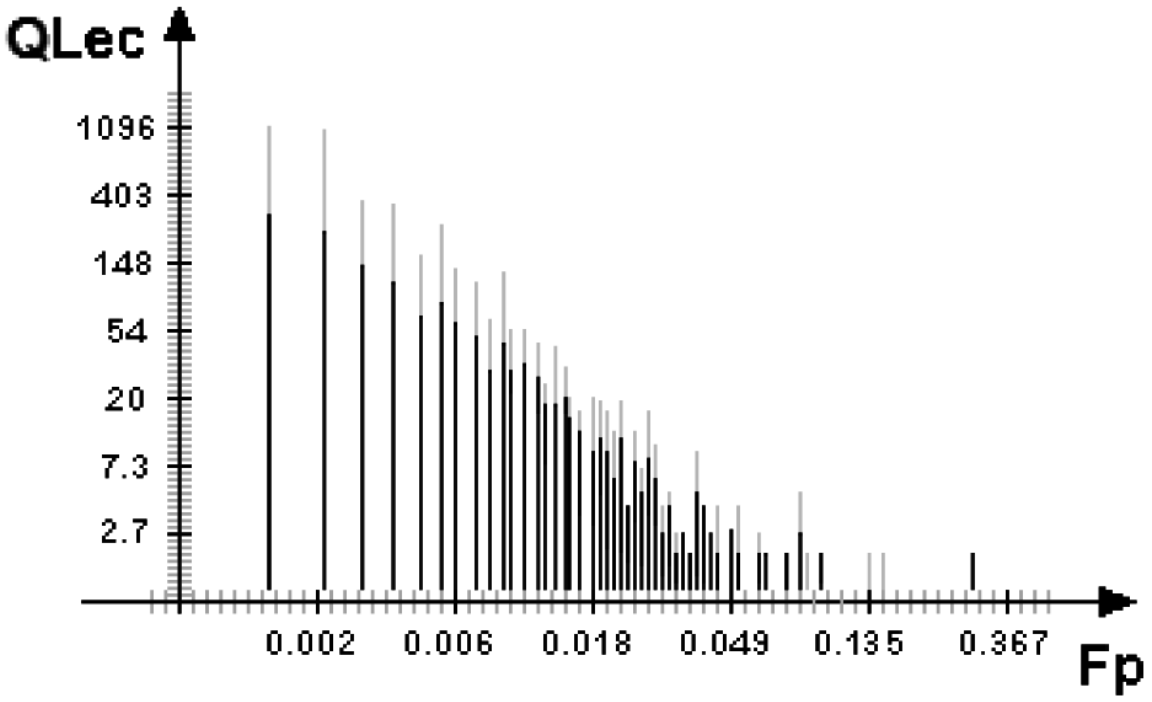

Рис. 3. Частотное распределение лексем алфавита поля «Наименование» БТИЗ после нормализации.

В ходе нормализации более длинные лексемы поглотили более короткие (Табл. 1).

Табл. 1. Количественное сокращение лексем алфавита поиска с учётом их длин.

длина лексемы (количество символов)	1	2	3	4	5	6	7	8	9	10
до нормализации	36	416	934	1283	1413	1374	1224	942	617	332
после нормализации	36	218	297	283	249	246	243	174	177	332
уменьшилась на	0%	47%	68%	78%	82%	82%	80%	81%	71%	0%

Самые короткие лексемы длиной в 1 символ, хоть и входят в более длинные лексемы, но частота их вхождения в объекты сущности заметно отличается от них, благодаря чему они остались в алфавите поиска сущности.

Были проведены исследования нормализации алфавитов поиска для 178 сущностей. В среднем нормализация уменьшила количество лексем в поисковом алфавите в 5.2 раза. При этом максимальное уменьшение количества лексем составило 17.3 раза, а

DOI: 10.7256/2305-6061.2015.4.17813

минимальное 1.2 раза.

По предварительной оценке трудоёмкости вычисления корреляций

Трудоёмкость вычисления корреляции = количество лексем алфавита * количество объектов * размер выборки

видно, что объём алфавита влияет на трудоёмкость вычисления корреляции, т.е. на идентификацию сущности по частотных характеристикам данных её объектов. Это было подтверждено практикой, скорость вычислений после нормализации поисковых алфавитов увеличилась примерно в 5 раз.

Благодаря нормализации алфавита увеличилась точность вычисления корреляции: сократилось расхождение вычисления корреляции (разность между максимальной и минимальной корреляциями для различных выборок) в среднем на 0.02036 на каждую корреляцию для 57458 участвующих в эксперименте вычислений. Предположительно это связано с тем, что стало меньше более коротких лексем. Ведь чем длиннее лексема, тем меньше вероятность её встречи в данных не относящихся к поисковой сущности, что понижает вероятность ложного срабатывания. Это предположение нуждается в проверке при дальнейших исследованиях.

Предложенная в работе методика нормализации алфавита поиска для повышения качества идентификации сущностей по частотным характеристикам их данных, заключается в удалении лексем алфавита входящих в другие лексемы алфавита с аналогичной частотой повтора в объектах сущности.

Экспериментально (на примере 178 сущностей) доказано, что данная методика позволяет в среднем в 5 раз уменьшить объём алфавита поиска, что значительно повышает быстродействие идентификации сущностей по частотным характеристикам их данных.

Благодаря уменьшению количества более коротких лексем методика нормализации позволяет уменьшить ошибку распознавания, как показали эксперименты в среднем на 0.02036 на каждую идентификацию.

Библиография :

1. Мальшаков Г.В. Методика повышения интероперабельности прикладного программного обеспечения на основе частотного анализа данных // Электротехнические комплексы и системы управления.-2015.-№ 3.-С. 67-70.
2. Мальшаков Г.В. Исследование ошибок идентификации сущностей прикладного программного обеспечения, выполняемой на основе частотного анализа данных // Наукоёмкие технологии.-2015.-№ 10.-С. 24-28.
3. ГОСТ Р 55062-2012 "Информационные технологии. Системы промышленной автоматизации и их интеграция. Интероперабельность. Основные положения"

DOI: 10.7256/2305-6061.2015.4.17813

4. Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии: Учеб. Пособие. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с.
5. Хомоненко А. Д., Цыганков В. М., Мальцев М. Г. Базы данных: Учебник для высших учебных заведений / Под ред. проф. А. Д. Хомоненко.-6-е изд., доп.-СПб.: КОРОНА-Век, 2009.-736 с.
6. Системы управления базами данных и знаний: Справ. изд. / А.Н.Наумов, А.М.Вендров, В.К.Иванов и др.; Под. ред. А.Н.Наумова. – М.: Финансы и статистика, 1991. – 352 с.: ил.

References:

1. Mal'shakov G.V. Metodika povysheniya interoperabel'nosti prikladnogo programmnogo obespecheniya na osnove chastotnogo analiza dannykh // Elektrotekhnicheskie kompleksy i sistemy upravleniya.-2015.-№ 3.-S. 67-70.
2. Mal'shakov G.V. Issledovanie oshibok identifikatsii sushchnostei prikladnogo programmnogo obespecheniya, vypolnyaemoi na osnove chastotnogo analiza dannykh // Naukoemkie tekhnologii.-2015.-№ 10.-S. 24-28.
3. GOST R 55062-2012 "Informatsionnye tekhnologii. Sistemy promyshlennoi avtomatizatsii i ikh integratsiya. Interoperabel'nost'. Osnovnye polozheniya"
4. Bashmakov A.I., Bashmakov I.A. Intellektual'nye informatsionnye tekhnologii: Ucheb. Posobie. – M.: Izd-vo MGTU im. N.E. Baumana, 2005. – 304 s.
5. Khomonenko A. D., Tsygankov V. M., Mal'tsev M. G. Bazy dannykh: Uchebnik dlya vysshikh uchebnykh zavedenii / Pod red. prof. A. D. Khomonenko.-6-e izd., dop.-SPb.: KORONA-Vek, 2009.-736 s.
6. Sistemy upravleniya bazami dannykh i znanii: Sprav. izd. / A.N.Naumov, A.M.Vendrov, V.K.Ivanov i dr.; Pod. red. A.N.Naumova. – M.: Finansy i statistika, 1991. – 352 c.: il.

DOI: 10.7256/2305-6061.2015.4.17813