SHORT NOTES

MSC 68W32 DOI: 10.14529/jcem170206

DEVELOPMENT OF ADAPTIVE MODEL FOR RECOGNITION OF TEXT INQUIRIES

A. S. Ambrosova, South Ural State University, Chelyabinsk, Russian Federation, asa1794@mail.ru

The paper is devoted to development of adaptive model to recognize the text inquiries using nomenclature reference book. The main calculations on resolution of conflicts which occur at recognition of text inquiries are submitted. Algorithms for solving conflicts use nomenclature reference book and parameters of association rules. Using such method, we take into consideration the previous experience of recognition and adapt the model in accordance with the new results of recognition. Adaptive model to recognize the text inquiries can be applied to different areas. For example, pharmaceutical features and drugs can be recognized in medicine in such a way.

Keywords: adaptive model, text inquiry, nomenclature reference book.

Introduction

For many years, the interest to methods of knowledge discovery in databases steadily grows [1, 4]. As a result, there are a lot of problems connected with processing of big data arrays in order to find new regularities, obtain and detect new knowledge. One of the problems is to recognize the goods by their key features [5].

1. Problem statement

Initial data for development of adaptive model are the nomenclature reference book of drugs with pharmaceutical parameters and patterns for the pharmaceutical parameters used to recognize the names of drugs.

The problem is to develop a self-training system, that will be able to recognize text inquiries and identify suitable drug from the catalog for the client.

Mathematical Problem statement.

Consider a set of drugs:

$$Drug = \{Drug_1, \ldots, Drug_x\}, \tag{1}$$

where $Drug_i$ is i-th drug, $i = \overline{1, x}$; x is a number of all drugs.

Let $Card$ be a catalog of the used drugs:

$$Card = (card_{ID}, A_{m \times n}, Drug_i), \tag{2}$$

where m is a number of lines in the catalog; n is a number of features in the catalog; $card_{ID}$ is an unique identifier in the catalog; $A_{m\times n} = \{a_i\ j\}, i = \overline{1,m}, j = \overline{1,n}$ are values of features for i-th drug.

Consider a database

$$D = \{T_1, \ldots, T_{N_D}\}, \tag{3}$$

where T_j is a transaction, $j = \overline{1, N_D}$; $N_D = |D|$ is a database power.

The elements T_j are presented in the following way:

$$T_j = (T_{ID_j}, Text_j, item_j, waiting_{ID}, reality_{ID}), \tag{4}$$

where T_{ID_j} is an identifier of j-th transaction T_j; $Text_j$ is an inquiry of the user (i.e. transaction); $item_j = \{\ldots, (e_l, mark_l), \ldots\}$ is a set of heuristic rules (i.e. heuristic) in j-th transaction, $1 \le l \le p$; e_l is a number of the heuristic found in the current transaction; $mark_l$ is a value of heuristics e_l; $waiting_{ID}$ is an identifier of the drug $card_{ID}$ distinguished by the system; $reality_{ID}$ is an identifier of the drug $card_{ID}$ distinguished by the user.

Consider a set of patterns which are used to recognize the current inquiry:

$$Pattern = \{pat_1, \ldots, pat_k\}, \tag{5}$$

where k is the total number of different patterns.

For every pattern there is corresponding heuristic $E = \{e_1, \ldots, e_p\}$, that defines the value for only one feature $Feature = \{f_1, \ldots, f_n\}$.

Several patterns can correspond to one particular heuristic, and several heuristics can correspond to one feature.

Each heuristic e_i has two features: S_{e_i} and C_{e_i} [2, 3], which are used to estimate the quality of recognition in the text, where S_{e_i} is a support of heuristic, C_{e_i} is an confidence of heuristic.

Each heuristic e_i and sets of heuristics [6], which are found by recognition of $item_j$, and their features are stored in the database of heuristics:

$$D_e = \{set_1, \ldots, set_p, set_{p+1}, \ldots\}, \tag{6}$$

where $set_i = \{< \ldots, e_l, \ldots >, S, C\}, i > p+1$ are sets of heuristics found by recognition, S is a support of the whole set of heuristics, C is an confidence of the whole set of heuristics.

For inquiry $s_p, p = N_{D+1}$, we need to recognize the features and choose suitable drug $Drug_i$ for user according to the catalog. Set up the model using the process of training.

2. Solving the problem

During the recognition of feature values, the following conflict situations between heuristics can take place in the text:

1. intersection of the set of heuristics and the catalog is empty;
2. two or more heuristics can detect different values of features by the same part of text;
3. two or more heuristics detect different values of the same feature in the text, besides parts of the text where corresponding features were found can be not intersected.

Using experimental data, we develop algorithms for resolving the conflict situations which are described above.

Algorithm to choose heuristic which belong to the considered part of the text.

1. Generate a set of all found unconflicted heuristics and add all possible sets of conflict values to the set (except $\varnothing$ and the whole conflict set).

2. In the generated set, choose 10% of all generated sets having the best average value of features $s * c$. If there is no such set in the reference book, then find s, c as the minimal established value, i.e. $(c := c_0)$.

Algorithm to choose a value of the feature among ones found by heuristics and eliminate an empty set in intersection of the set of heuristics and the catalog.

1. If the same heuristic gives different values of the same feature, then choose the first generated set.

2. If different heuristics detect the same value of the same feature, then generate a set containing all found heuristics.

3. Generate all possible sets of values of the first N significant features.

4. Choose sets among ones generated at the previous step, which are mentioned in the catalog. If there are no such sets, reduce the number of significant features, and go to the previous step.

5. Choose 10% of all found sets having the best average value of features $s * c$. Record the first N features for each remained set.

6. For each of the remained features, generate all possible values and add to the existing recorded set.

7. Choose the sets among them which are mentioned in the reference book.

8. Among the found sets, choose sets with the best average value of features $s * c$. If there are no such sets, then delete the value of this feature.

9. Add the found value of feature to the recorded set and go to the following feature.

Algorithm of adaptive model to recognize text inquiries.

1. Require the user to enter the data as a text inquiry string s_p. Connect additional information: the catalog of the used drugs $Card$; the features set $Feature$; the heuristics set E; the patterns set $Pattern$.

2. Consistently apply each pattern pat_i to an initial inquiry string. For each pattern, determine the interval $[a_{pat_i}; b_{pat_i}]$ of covering text. Correspond a heuristic e_k to every obtained pattern pat_i and define a value of corresponding feature f_l.

3. Detect and resolve all conflicts connected with intersections of heuristics by the text.

4. Detect and resolve all conflicts connected with intersections by the values of features and an empty intersection of heuristics set and the catalog.

5. Output the result of recognition, i.e. the suitable drug $Drug_i$ that corresponds to current inquiry and is contained in the catalog $Card$.

6. Require the user to confirm or reject the result of program recognition.

7. Update support S and confidence C for the set of heuristics and for all its subsets found during recognition process.

Conclusion

We consider and resolve main types of the conflicts which appear during the text recognition process using heuristic rules. The algorithm of adaptive model for recognition of the text inquiries is developed.

References

1. Bay Vo, Bac Le, Fast Algorithm for Mining Generalized Association Rules. *International Journal of Database Theory and Application*, 2009, vol. 2, no. 3, pp. 1–12. Available at: http://www.sersc.org/journals/IJDTA/vol2_no3/1.pdf (accessed on 20 June 2017)

2. Sebastiani F. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 2002, vol. 34, no. 1, pp. 1–47. Available at: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf (accessed on 20 June 2017)

3. Zayko T.A., Oleynik A.A., Subbotin S.A. Associative Rules in Data Mining. *Vestnik NTU "KhPI"*, 2013, no. 39 (1012), pp. 82–96. (in Russian)

4. Chubukova I.A. *Data Mining.* Moscow, Internet – Un-t Inform. Tekhnologiy: BINOM. Lab. znaniy, 2008. (in Russian)

5. Sidorova E.A., Anohin S.V., Kononenko I.S., Salomatina N.V. Thematic Analysis of User Queries Based on Subject Dictionaries. *Vestnik Novosibirskogo Gosudarstvennogo Universiteta. Seriya: Informatsionnyye Tekhnologii*, 2014, vol. 12, no. 4, pp. 83–94. (in Russian)

6. Dobrov B.V., Lukachevich N.V. Automatic Rubrication of Full Text Documents on Qualifiers of Complex Structure. [*VIII Nats. Konf. po Iskusstvennomu Intellektu KII-2002 – VIII National Conference of Artificial Intelligence KII-2002*]. Moscow, Fizmatlit Publ., 2002, vol. 1, pp. 178–186. (in Russian)

Anastasya S. Ambrosova, Under Graduate, Department of Applied Mathematics and Programming, South Ural State University (Chelyabinsk, Russian Federation), asa1794@mail.ru

Received May 15, 2017

УДК 004.89 **DOI: 10.14529/jcem170206**

РАЗРАБОТКА АДАПТИВНОЙ МОДЕЛИ РАСПОЗНАВАНИЯ ТЕКСТОВЫХ ЗАПРОСОВ

А. С. Амбросова

Работа посвящена разработке адаптивной модели распознавания текстовых запросов с использованием номенклатурного справочника. Представлены основные выкладки по разрешению конфликтов, возникающих при распознавании текстовых запросов. Разработанные алгоритмы, разрешают конфликтные ситуации, руководствуясь соответствием с номенклатурным справочником и выбирая эвристические правила с лучшими показателями характеристик. Такой подход позволяет учитывать предыдущий опыт распознавания текстовых запросов и адаптировать модель в соответствии с полученными результатами распознавания. Адаптивная модель распознавания текстовых запросов может использоваться в различных областях, например, в медицине для выявления фармацевтических параметров и лекарств.

Ключевые слова: адаптивная модель, текстовый запрос, номенклатурный справочник.

Литература

1. Bay, Vo. Fast Algorithm for Mining Generalized Association Rules / Bay Vo, Bac Le // International Journal of Database Theory and Application. – 2009. – V. 2, № 3. – P. 1–12. Доступ: http://www.sersc.org/journals/IJDTA/vol2_no3/1.pdf (запрос 20 июня 2017)
2. Sebastiani, F. Machine learning in automated text categorization / F. Sebastiani // ACM Computing Surveys. – 2002. – V. 34, № 1. – P. 1–47. Доступ: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf (запрос 20 июня 2017)
3. Зайко, Т.А. Ассоциативные правила в интеллектуальном анализе данных / Т.А. Зайко, А.А. Олейник, С.А. Субботин // Вiсник НТУ «ХПI». Серiя: Iнформатика та моделювання. – 2013. – № 39 (1012). – С. 82–96.
4. Чубукова, И.А. Data Mining / И.А. Чубукова. – М.: БИНОМ. Лаборатория знаний. – 2008.
5. Сидорова, Е.А. Тематический анализ запросов пользователей на основе предметно-ориентированного словаря / Е.А. Сидорова, С.В. Анохин, И.С. Кононенко, Н.В. Саломатина // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2014. – Т. 12, № 4. – С. 83–94.
6. Добров, Б. В. Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры / Б.В. Добров, Н.В. Лукашевич // VIII Нац. конф. по искусственному интеллекту КИИ-2002. – М.: Физматлит, 2002. – Т. 1. – С. 178–186.

Амбросова Анастасия Сергеевна, бакалавр, кафедра прикладной математики и программирования, Южно-Уральский государственный университет (г. Челябинск, Российская Федерация), asa1794@mail.ru

Поступила в редакцию 15 мая 2017 г.