MODEL AND ALGORITHM FOR ADAPTIVE SEARCH BY LOGICAL EXPRESSIONS

Savenko A. G., Sherstnev A. S.
Institute of Information Technologies of the Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus
E-mail: savenko@bsuir.by

The article presents the developed graph model and algorithm for adaptive search by logical expressions using the example of a web application for the selection of employees of IT-companies based on their competencies. The analysis of the performance of the search algorithm by logical expressions for borderline cases is carried out. The flexibility of the search algorithm is shown when building queries by minimizing the logical expression of a search query.

Introduction

Information retrieval is an essential part of any automated system, including information systems related to the processing of text data. The important criteria for evaluating the performance of search algorithms are their speed and relevance. This article proposes a general model of a search algorithm that will allow you to efficiently and flexibly, from the point of view of search queries, determine the subset of data of interest to the user.

I. Data storage organization model

From the point of view of the information model, the proposed search engine implements a complex search based on logical expressions.The information database is stored in the form of relationships between entities, the combination of which allows you to achieve the desired search criterion.Entities should describe as little information as possible in order to provide less granularity and therefore provide more accurate results. Links should be one-way and directional, and their number should be minimal.The described data storage structure can be easily implemented as a graph database. An illustration of a graph database model using the example of an information system for the selection of employees of an IT-company to work on project tasks [1] in accordance with their competencies is shown in figure 1.

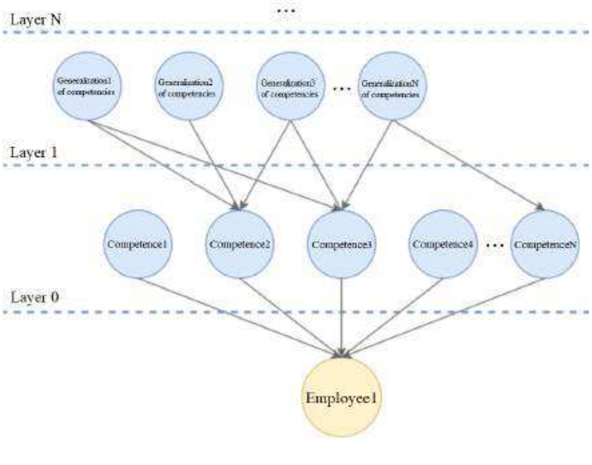
...
Layer N
Generalization1 of competencies
Generalization2 of competencies
Generalization3 of competencies
...
GeneralizationN of competencies
Layer 1
Competence1
Competence2
Competence3
Competence4
...
CompetenceN
Layer 0
Employee1

Рис. 1 – Sample graph database model

As you can see from the sample database model, the data is organized in layers. At the lowest level is the employee information node. This node is linked to skill nodes by one-way communication. In turn, skill nodes are associated with the next layer of generalizations of these skills into broader groups of concepts. This generalization will allow the search system to find employees according to broader concepts of their competencies. Also, on one conditional layer, links between nodes are prohibited to prevent cyclic search.

II. Search algorithm

The adaptive search in the proposed system is carried out by logical expressions. A search query can include basic logical operations such as AND, OR, as well as a grouping operation with priority, and the search process itself is a graph traversal. However, first, it is necessary to transform the input expression into a disjunctive normal form (DNF), where in the given example, the competencies of employees act as logical literals. DNF will allow converting any incoming logical expression into a disjunction of conjunctions of literals, which will make it possible to split the algorithm into two stages:

1. Search for all employees with several competencies at the same time:
   - the first step is to convert the original expression to DNF;
   - the next steps are a cycle through all conjunctive groups and the formation of queries to traverse the graph, taking into account the inclusion of all competencies from the group.
2. Combining all results from the previous step and removing duplicates.

DNF will also allow minimizing a logical expression, which will speed up the search algorithm by reducing the number of logical literals. Two-way breadth first search is used to find paths on a graph.

III. Search algorithm efficiency analysis

An important criterion in the development of an adaptive search algorithm is its performance (since it is necessary to process a large amount of data in an acceptable user waiting time), as well as flexibility in building search queries. Based on this, we will consider the boundary (worst) cases of the algorithm operation, in which the search will be as ineffective as possible. The number of connections between the nodes of two layers at a given level is determined by the function $L(k, m, n)$, where k is the number of nodes in the first layer, m is the number of nodes in the second layer, n is the number of the level, and at the same time $k > 0; m > 0; n \geq 0; k, m, n \in N$. The number of nodes in the next layer is set by the function $G(t, n)$, where t is the number of nodes of the previous layer, n is the number of the level, and at the same time $t > 0; n \geq 0; t, n \in N$. The total number of links in a multilayer graph, with restrictions on links, is described by formula 1. The total number of nodes is determined by formula 2. The extreme worst case for a search would be the number of links that needs to be checked from the topmost layer to the bottommost, that is, traverse the entire graph. The function describing this extreme case creates connections between each node of the initial layer and each node of the next one and has the form $L_{max} = (k, m, n) = km$ Then the growth rate of the number of links $R(1, n, L_{max}, G_l)$ and the number of nodes $E(1, n, G_l)$ at n levels will have the form shown in Figures 2 and 3.

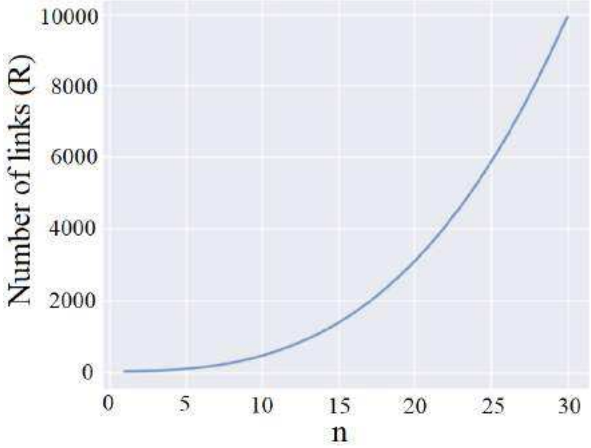

Рис. 2 – Graph of the growth rate of the number of links for n number of layers

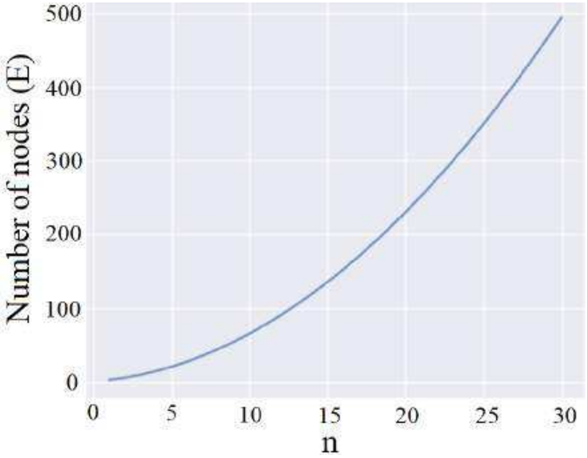

Рис. 3 – Graph of the growth rate of the number of nodes for n number of layers

Obviously, even with a small number of layers $n \geq 30$, the number of nodes and links is quite different in order, and the dependence itself is far from linear and closer to a power-law or exponential.

IV. Conclusions

As a result of the study, a model of information system data organization and a flexible adaptive search algorithm by logical expressions were developed. The data organization model is a graph database. The created algorithm has a high relevance of the results obtained and a high speed of obtaining them. The developed model and search algorithm are applied in the implemented cross-platform system for recruiting employees of IT-companies based on their competencies.

V. Bibliography

1. Savenko, A. G. Web system for the adaptive search for employees of IT-companies based on their competencies by logical expressions / A. G. Savenko, A. S. Sherstnev // Web programming and Internet technologies WebConf2021: materials of the 5th International scientific and practical conference –Minsk, BSU –2021. – P. 127–129.

$$R(t, n, L, G) = \begin{cases} (R(G(t, n), n-1, L, G) + L(t, G(t, n), n), n > 0 \\ 0, n = 0 \end{cases} , \quad (1)$$

where t is the number of nodes on the initial layer; $t > 0; t \in N$; n is the total number of layers; $n \geq 0; n \in N$; L is a function that determines the number of connections between the nodes of two layers at a given level; G is a function that determines the number of nodes in the next layer.

$$E(t, n, G) = \begin{cases} E(G(t, n), n-1, G) + G(t, n), n > 0 \\ 0, n = 0 \end{cases} , \quad (2)$$

where t is the number of nodes on the initial layer; $t > 0; t \in N$; n is the total number of layers; $n \geq 0; n \in N$; G is a function that determines the number of nodes in the next layer.