**УДК 681.3**

***Р. Даревич, *Д. Досин, В. Литвин**
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж,
*Фізико-механічний інститут ім. Карпенка

# МЕТОД ПОБУДОВИ ІНТЕЛЕКТУАЛЬНИХ МЕТАПОШУКОВИХ СИСТЕМ НА ОСНОВІ АДАПТАЦІЇ ОНТОЛОГІЇ



**Наведено метод побудови інтелектуальних метапошукових систем на основі адаптації онтології до потреб користувачів. Для здійснення адаптації вводяться ваги понять онтології та ваги зв'язків між ними. Автори запропонували метрику, за допомогою якої визначається відстань між текстовими документами. Використовуючи регресійний аналіз, метрика переводиться в числову оцінку рангування знайдених документів.**

**This article considers intelligent metasearch systems based on ontology adaptation for users. Ontology term and connection weights are proposed for the ontology adaptation. Authors propose metric that helps to define distance between text documents. Using regressive analysis metric is transformed into document rank.**

## Вступ

### 1. Постановка проблеми

Швидкий розвиток галузі інформаційного пошуку пов'язаний із появою та розбудовою глобальної комп'ютерної мережі Інтернет, яка створила принципово нові умови та можливості застосування інформаційних технологій для доступу, пошуку, опрацювання та зберігання інформації. За таких обставин для ефективного пошуку потрібної (релевантної) інформації необхідні автоматизовані інформаційно-пошукові системи, які ґрунтуються на інтелектуальних алгоритмах аналізу текстів. Аналіз існуючих підходів створення високоефективних технологій автоматизації інформаційного пошуку текстових документів засвідчив переваги адаптивних інтелектуальних метапошукових систем (МПС). Оскільки робота таких систем не передбачає постійної взаємодії з користувачами, якість пошуку визначається точністю подання їх інформаційних потреб, що визначаються предметною областю (ПрО) користувача.

### 2. Аналіз останніх досліджень та публікацій

Статистичні та семантичні методи пошуку, відповідно до способів подання інформаційних потреб (векторно-просторова модель, міра на основі коефіцієнта Дайса, латентно-семантичне індексування, порівняння концептуальних графів) запропонували С. Думайс, Дж. Солтон, Е. Расмусен та інші. Загальним недоліком цих методів є недостатня точність порівняння документів за змістом. Водночас для автоматизованих систем інформаційного пошуку, в яких не передбачено інтерактивної взаємодії системи з користувачем, така точність має вирішальне значення.

Одним із способів підвищення точності порівняння документів за змістом є використання в складі МПС онтології – множини понять ПрО, пов'язаних семантичними зв'язками та визначеними для них функціями інтерпретації. Сьогодні розроблено низку таких методів (М. Монтес-Гомез, Ванг Гуі-джин, Г. Бульсков, Д. П. Ночевнов), проте в них онтологія є статичною, вагові коефіцієнти понять призначаються вручну, що утруднює їхнє ефективне застосування в автоматизованих МПС. Вирішити цю проблему можна, використовуючи в алгоритмі роботи системи процедури адаптації її онтології до заданої ПрО та інформаційних потреб користувача. Методи ж автоматичної адаптації онтології, які не передбачають безпосередньої участі користувача, сьогодні розвинуті недостатньо, що значно обмежує використання адаптивних онтологій в МПС. Тому розроблення методів та алгоритмів адаптації онтології автоматизованої МПС під час її експлуатації до інформаційних

потреб користувачів становить актуальну наукову задачу, розв'язання якої сприятиме підвищенню ефективності інформаційного пошуку, а також зменшенню часових і фінансових затрат на створення та обслуговування таких систем.

## 3. Формулювання цілі статті

*Мета роботи* – підвищити точність оцінювання подібності текстових документів за їх змістом у МПС шляхом розроблення методів та алгоритмів адаптації онтології до інформаційних потреб користувачів на основі вдосконалення методу визначення коефіцієнтів важливості понять та зв'язків між ними.

## 4. Виклад основного матеріалу

Адаптивною вважаємо онтологію, здатну налаштовуватись на певну ПрО шляхом зміни своєї структури і значень параметрів. Серед властивостей адаптивної онтології ключовою є її здатність під час експлуатації інтелектуальної МПС динамічно формуватися, що зумовлює необхідність періодичної оптимізації структури та змісту такої онтології. Під час створення МПС до ядра її онтології вносяться базові поняття, семантичні зв'язки між ними, механізми наповнення і оптимізації. Побудова онтології можлива з різною мірою автоматизації: вручну за допомогою інженера зі знань, напівавтоматично – використовуючи діалогові програми чи спеціалізовані редактори онтологій, або ж автоматично – видобуваючи знання методами інтелектуального аналізу текстових документів.

Огляд літератури підтвердив існування труднощів із створенням адаптивних онтологій, придатних для промислової експлуатації у складі МПС. Для формування онтологій, як правило, використовують засоби ручного та інтерактивного напівавтоматичного наповнення, що зумовлює значні фінансові та часові затрати, переважно не сумісні з комерційним застосуванням таких систем. Показано, що автоматичне наповнення онтології шляхом видобування знань з природомовних текстів та використання процедур оптимізації її структури та змісту підвищує ефективність роботи МПС внаслідок їх налаштування на ПрО користувача. Проте, будуючи такі системи, необхідно враховувати їх швидкодію, обмеження на максимальний обсяг доступної робочої пам'яті та можливість виникнення логічних конфліктів між даними, отриманими від різних джерел.

Проведений аналіз переваг та недоліків існуючих моделей подання знань показав, що для моделювання структури та функцій онтології МПС необхідно використати поєднання різних моделей подання знань: фреймової – для опису загальної таксономічної структури ПрО, мережевої (концептуальних графів) – для відображення існуючих у цій ПрО семантичних зв'язків між окремими поняттями та їх властивостями, логіки предикатів та правил продукцій – для реалізації механізмів міркування, контролю цілісності, наповнення та оптимізації структури та змісту онтології.

Одним з підходів до реалізації механізмів оптимізації є автоматичне зважування понять онтології та семантичних зв'язків між ними під час експлуатації системи. Цю роль виконують коефіцієнти важливості понять та зв'язків, означені як числова міра, котра характеризує значимість даного поняття (зв'язку) у конкретній ПрО і змінюється за визначеним алгоритмом (правилами) під час опрацювання текстових документів. Розподіл коефіцієнтів має відповідати таким основним вимогам:

відображати семантичну вагу понять ПрО, в якій ця інтелектуальна система реально застосовуватиметься;

формуватися під час наповнення онтології та коректуватись за визначеним алгоритмом;

забезпечувати контроль цілісності онтології;

задовольняти вимоги метрики під час їх використання для порівняння семантичної близькості понять.

Модель адаптивної онтології подамо у вигляді п'ятірки: $G(C,R,F,W,L)$, де $C$ – скінченна множина атомарних понять ПрО; $R$ – скінченна множина семантичних зв'язків між атомарними поняттями ПрО; $F$ – скінченна множина функцій інтерпретації, яка встановлює аксіоматичну взаємозалежність понять з множини $C$ через множину зв'язків $R$; $W$, $L$ – множина коефіцієнтів важливості понять та зв'язків відповідно, обчислювати які запропоновано за таким алгоритмом:

1. Повна вага $W_j^i$ класу онтології дорівнює сумі власної ваги $Wo_j^i$, ваги підкласів $Ws_j^i$ та ваги суміжних класів $Wn_j^i$ (класів, пов'язаних з цим класом не „is-a” зв'язком):

$$W_j^i = Wo_j^i + Ws_j^i + Wn_j^i, \quad (1)$$

де $Ws_j^i = \sum_k Wc_k^{i+1} \cdot L_{j,k}$ − вага $k$ підкласів $j$-го класу $i$-го рівня;

$Wc_k^{i+1} = Wo_k^{i+1} + Ws_k^{i+1}$ − вага класу $C_k^{i+1}$; $L_{j,k}$ − вага зв’язку між класами $C_j^i$ та $C_k^{i+1}$.

2. У момент внесення на *i+1-й* рівень нового підкласу йому присвоюється власна вага $Wo_j^{i+1}$, що дорівнює половині власної ваги класу вищого ($i$-го) рівня. Вага класу $Wc_j^i$ та усіх батьківських класів аж до кореневого збільшується на величину ваги новоствореного підкласу:

$$Wc_j^m = Wc_j^m + Wo_j^{i+1}, \forall m \le i. \quad (2)$$

3. Під час встановлення зв’язку між поняттями $k_1$ та $k_2$ між відповідними вершинами графу онтології з’являється ребро, а до ваги суміжних класів $Wn_1$ додається вага $Wc_2$ і, навпаки, до $Wn_2$ додається вага нового, суміжного до нього, класу $Wc_1$:

$$Wn_j = \sum_k Wc_k \cdot L_{j,k}. \quad (3)$$

4. Вага екземпляра у базі знань дорівнює повній вазі (1) його класу в онтології.

Розроблений алгоритм покладено в основу методу автоматичного перерахунку ваги класів онтології та екземплярів бази знань під час її наповнення та налаштування на задану ПрО користувача під час експлуатації.

Засобами Delphi реалізовано імітаційну модель процесу генерування та оптимізації онтології (рис. 1) для дослідження ефектів, пов’язаних з обчисленням вагових коефіцієнтів.

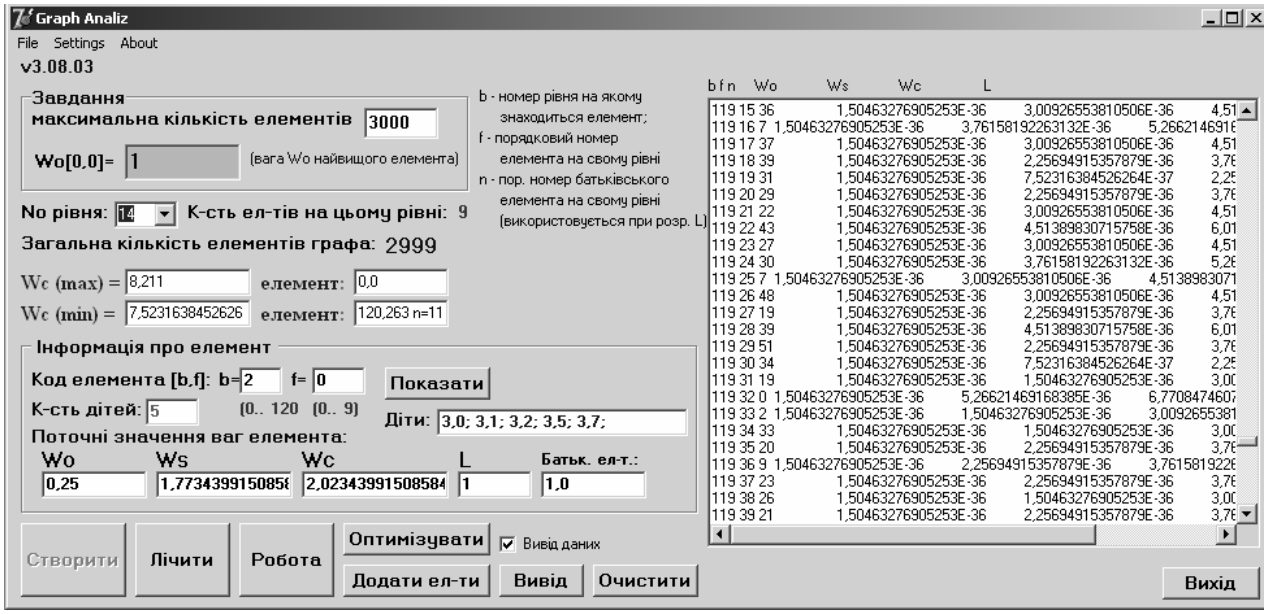


*Рис. 1. Програмний інтерфейс імітаційної моделі адаптивної онтології*

Зокрема, дослідження стосувалося:

виявлення можливих зворотних залежностей (циклів) під час обчислення коефіцієнтів важливості елементів онтології;

оцінювання кількості елементів з однаковою мінімальною вагою;

визначення діапазону значень, яких можуть набувати коефіцієнти важливості понять.

У моделі застосовано статистику розподілу елементів за рівнями таксономії лексичної бази даних WordNet.

За результатами моделювання процесу генерування структури онтології встановлено, що результуюче відношення між вагою понять, близьких до кореневого, та вагою понять нижніх рівнів становить кілька порядків (рис. 2), тому вилучення чи внесення до онтології бази знань понять нижніх рівнів під час оптимізації не змінює відчутно вагу решти понять, що допускає зведення задачі оптимізації змісту до задачі лінійного програмування.

Досліджено розподіл елементів онтології за діапазонами ваг для визначення кількості елементів з мінімальною однаковою вагою (рис. 3). Їхня кількість повинна бути обмежена для того, щоб при подальшому вилученні цих елементів як найменш цінних з погляду їх важливості у цій ПрО не виникала неоднозначність вибору. Діапазони ваг вибрано шляхом поділу різниці між максимальною та мінімальною вагою елементів на однакові частини. Встановлено, що розподіл дає змогу однозначно вибирати в онтології 10% елементів з мінімальною вагою від їхньої загальної кількості.

Отримані результати дослідження методу підтвердили можливість його застосування для розроблення алгоритмів оптимізації онтології з урахуванням цінності інформації, що в ній міститься.

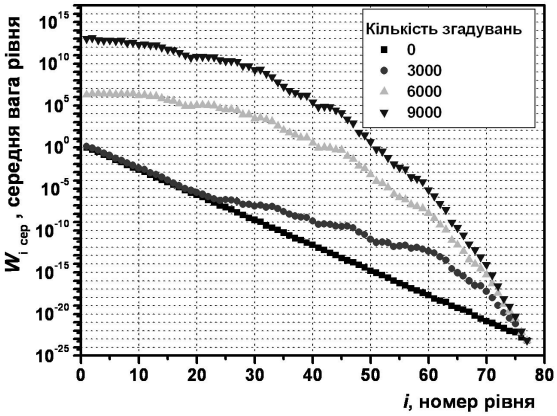


*Рис. 2. Зміна середньої ваги на рівнях графу залежно від частоти згадування зв'язків в онтології*

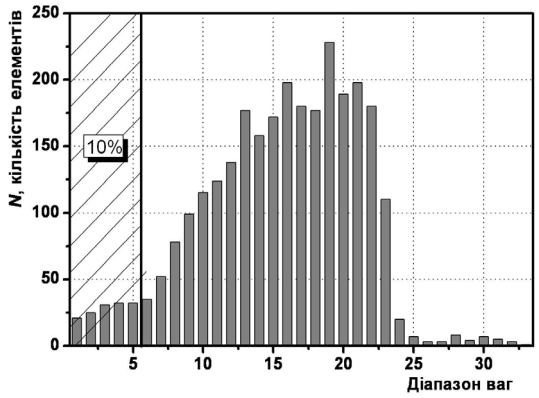


*Рис. 3. Розподіл кількості елементів онтології за діапазонами ваг*

Нами розроблено метод оцінювання подібності за змістом електронних текстових документів, який полягає у поданні текстів у вигляді концептуальних графів, доповненні їх відповідним контекстом та коефіцієнтами важливості з адаптивної онтології, знаходженні вершин, що є центрами семантичної ваги таких графів, та обчисленні семантичної відстані між знайденими центрами.

Згідно з розробленим методом, семантичну відстань між двома вершинами $C_i$ та $C_j$ графу, якщо вони з'єднані дугою, запропоновано визначати виразом:

$$d_{ij} = \frac{Q}{L_{ij}(W_i + W_j)}, \tag{4}$$

де добуток $L_{ij}\left(W_i + W_j\right)$ – сила зв'язку між вершинами $C_i$ та $C_j$, $Q$ – константа, яка залежить від конкретної онтології. За замовчуванням прийнято, що $L_{ii} = \infty$, тоді $d_{ii} = 0$. Для обчислення найкоротшого шляху $d_{ij}^*$ між вершинами $C_i$ та $C_j$ застосовано відомий алгоритм Дейкстри.

Для визначення центру ваг концептуального графу (вершини $C_{i^*}$) необхідно знайти мінімальну середню відстань $\overline{d}_{i^*} = \min_i \overline{d}_i$, де середня відстань $\overline{d}_i$ для кожної вершини $C_i$ обчислюється за формулою:

$$\overline{d}_i = \frac{\sum_{j=1, j\neq i}^{n} d_{ij}^*}{n-1}, \tag{5}$$

де $n$ – кількість вершин графу.

Визначені так центри ваг використано для знаходження відстані між концептуальними графами двох порівнюваних документів.

Після цього, накладаючи отримані графи з визначеними їх центрами, отримано суміщений граф. При цьому вага спільних вершин у кінцевому графі визначається як середнє арифметичне ваг цих вершин у відповідних графах до накладання. Вага ж вершин, що не є спільними для цих графів, у суміщеному графі залишається незмінною. У випадку, якщо порівнювані графи не мають спільних вершин, відстань між ними прийнято рівною ∞, тому відповідні тексти не є подібними.

Якщо побудовано суміщений граф, в якому $C^1$ – центр ваги першого графу, а $C^2$ – другого, то визначається мінімальна відстань $d^{12}$ між цими центрами:

$$d^{12} = \min d(C^1, C^2). \qquad (6)$$

Отримана відстань дає оцінку подібності змісту двох текстів, яким відповідають ці концептуальні графи. Чим ця відстань є меншою, тим подібніші є порівнювані тексти. У роботі нами було показано, що такий метод порівняння змісту природомовних текстів задовольняє усі три аксіоми метрики.

На основі вибраних та обґрунтованих критеріїв оптимальності онтології МПС нами розроблено метод оптимізації її структури та змісту.

Критеріями оптимальності є: фізичний обсяг пам'яті, швидкодія, повнота онтології, її цілісність та збалансованість, причому критерій цілісності застосовується в процедурах нормалізації структури, тобто мінімізації надлишковості та усунення логічних суперечностей. Реалізація процедур оптимізації змісту відбувається за критеріями: обмеження на фізичний об'єм, повнота та швидкодія.

Автоматичне генерування онтології зумовлює необхідність здійснювати її локальну оптимізацію під час наповнення, і глобальну – на етапі впорядкування, коли процес наповнення призупинено до завершення процедури оптимізації. Метод оптимізації онтології містить задачу нормалізації її структури та задачу оптимізацію змісту. При цьому нормалізація передбачає виявлення та усунення паралельних ребер, циклів, петель, дублювання вершин з аналогічними параметрами та інших особливостей структури графу онтології, які порушують її цілісність та знижують ефективність функціонування. Для збільшення інформаційної насиченості онтології виконується процедура оптимізації її змістової частини, яка полягає у визначенні та вилученні заданої частки найменш важливої для користувача МПС інформації. З метою збереження цілісності онтології спершу перевіряють її структурну узгодженість, а потім вибирають найважливіші поняття, які становлять решту істинних тверджень. Процедуру оптимізації змістової частини онтології доцільно здійснювати шляхом послідовної редукції її графу до задоволення вимог установлених критеріїв оптимальності.

*Задача нормалізації структури графу онтології* складається з двох підзадач: усунення надлишковості та усунення суперечностей. За подання структури онтології зваженим графом, де вага ребра відображає важливість представленого ним зв'язку та визначається через частоту його вживання, а надлишковість проявляється у вигляді паралельних ребер та петель, усунення цих та інших особливостей полягає у послідовному вилученні ребер з мінімальною вагою зі збереженням зв'язності усього графу. Цю задачу розв'язано шляхом застосування процедури виділення мінімального остова.

Автоматизоване внесення до онтології нових тверджень зумовлює виникнення внутрішніх логічних конфліктів, що порушує її цілісність, тому систему треба забезпечити здатністю виявляти та вилучати їх. Розроблено відповідний алгоритм, який ґрунтується на застосуванні методу резолюцій. Він складається з таких кроків:

1) знання подаються в логічній формі;

2) правильно побудовані формули числення предикатів спрощуються до вигляду речень шляхом виконання стандартних операцій;

3) застосовується метод резолюцій для виявлення суперечностей;

4) з суперечливих тверджень вилучається те, в якого коефіцієнт достовірності джерела менший.

Достовірність джерела твердження означено як імовірність отримання від нього істинного твердження $D_n = P(s=1)$. Вважають, що апріорна достовірність незнайомого джерела дорівнює 0,5.

Апостеріорну достовірність під час поступової перевірки істинності $s$ наданих $n$-м джерелом тверджень визначають за формулою:

$$D_{n,i+1} = \frac{D_{n,i}}{(2-s)} + \frac{1-D_{n,i}}{2} \cdot s, \tag{7}$$

де $s$ – істинність твердження, що набуває значення 1, якщо твердження істинне, або 0 – у протилежному випадку, $i$ – номер кроку підтвердження/заперечення істинності одного з тверджень $n$-го джерела.

Для розв'язання *задачі оптимізації змісту онтології МПС* необхідно на основі критеріїв швидкодії та повноти визначити оптимальну кількість понять такої онтології. Для заданих критеріїв цільова функція:

$$f(K) = \overline{E} + \frac{1}{G} \to \min, \tag{8}$$

$$\overline{E} = \frac{N\sum_{i=1}^{k}[P(i)\cdot(i+k)]}{K}, \quad K = N\sum_{i=1}^{k} P(i), \tag{9}$$

де $\overline{E}$ – швидкодія, виражена як середній ексцентриситет вершин графу, що представляє онтологію; $k$ – кількість рівнів у графі; $G$ – відносна кількість понять в онтології, $G = K/N$; $K$ – кількість понять в онтології; $N$ – кількість понять у словнику ПрО.

На основі аналізу лексичної бази даних *WordNet* досліджено статистику розподілу елементів за рівнями типової онтології $P(i)$. Встановлено, що за такого розподілу означених критеріїв оптимальності та словника, який містить 100000 понять, мінімум цільової функції (8) відповідає оптимальній кількості 31000 понять в адаптивній онтології МПС (рис. 4).

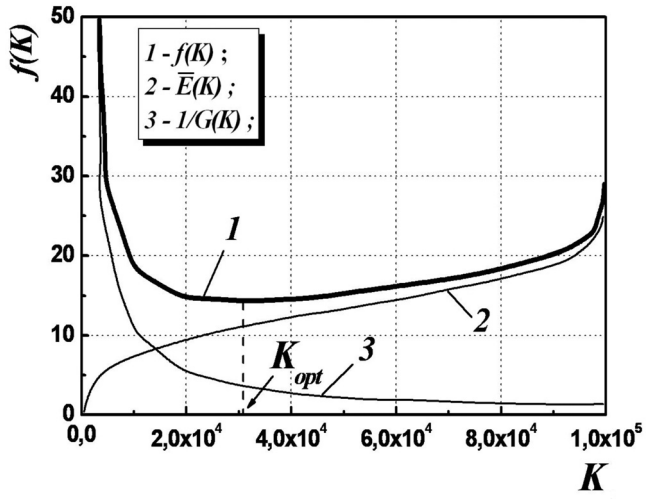


*Рис. 4. Цільова функція пошуку оптимальної кількості понять онтології МПС*

Під час наповнення онтології до визначених так меж виникає необхідність періодично вибирати і вилучати з онтології певний надлишковий об'єм даних з урахуванням коефіцієнтів важливості її елементів. За результатами виконаного моделювання процесу генерування та оптимізації онтології МПС процедура вилучення з онтології елементів з найменшою вагою може бути зведена до дискретної оптимізаційної задачі, а саме задачі про рюкзак.

Нехай онтологія складається з $n$ елементів загальним об'ємом пам'яті $M$. Роль „рюкзака" виконує певна задана частка обсягу $N = 1/10M$, до якої треба віднести найменш цінні елементи (поняття з мінімальною семантичною вагою та максимальним об'ємом) для подальшого їх вилучення. Тоді необхідно максимізувати сумарний зиск:

$\sum_{i=1}^{n} \frac{1}{W_i} x_i$ таких елементів *i*, для яких $\sum_{i=1}^{n} m_i x_i \le N$ та $W_i > 0$, $m_i > 0$, $i = 1,...n$,

де $x_i$ – поняття онтології, $x_i = 1$, якщо поняття вносимо в "рюкзак" та 0 – у протилежному випадку; $W_i$ – вага поняття; $m_i$ – об'єм пам'яті, який займає цей елемент.

Задача спрощується, якщо вважати, що об'єм робочої пам'яті, зайнятий *i*-м елементом, $m_i = m = const$, що, як правило, відповідає умовам реалізації онтології МПС Показано, що сформульовану задачу можна розв'язати за допомогою жадібного алгоритму.

Використовуючи імітаційну модель, описану в другому розділі, експериментально показано, що через 40–50 циклів роботи (навчання, робота, оптимізація) відносна кількість видалених елементів, що були додані в попередньому циклі роботи, сягає ~ 65 % і майже не змінюється (рис. 5). Це означає, що критична маса важливих для даної ПрО понять внесена в онтологію. На основі цього можна стверджувати, що онтологія налаштована на задану ПрО, тобто адаптована до інформаційних потреб конкретного користувача.

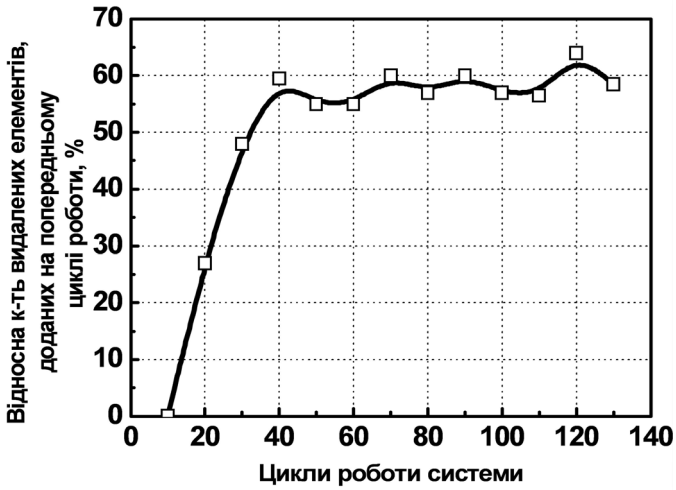


*Рис. 5. Зміна відносної кількості елементів онтології, доданих на попередньому циклі роботи залежно від кількості циклів*

Нами розроблено архітектуру МПС у складі віртуального автоматизованого робочого місця (ВАРМ), призначення якої – пошук електронних текстових документів у мережі Інтернет в автономному режимі за документом-взірцем. Функціонування системи забезпечується динамічним наповненням її онтології, що супроводжується оптимізацією, методи якої розроблені у дисертації.

Служба автоматично виділяє з усього доступного масиву електронних текстових документів підмножину релевантних до документа-взірця, визначеного користувачем. Особливістю МПС є наявність у її складі адаптивної онтології, здатної відображати інформаційні потреби користувача, що забезпечує автономний (без його участі) режим пошуку наукових публікацій (моніторингу нових надходжень).

Основними компонентами розробленої МПС (рис. 6) є:

пошуковий агент, реалізований на базі Wget, який працює під управлінням ОС Linux і забезпечує видобування анотованих публікацій з мережі Інтернет;

супровідна база даних під управлінням СУБД MySQL, в якій зберігаються профілі користувачів, відповідні ПрО, запити користувачів, а також знайдені анотації;

програмний пакет синтаксично-семантичного аналізу на базі Link Parser, який забезпечує побудову семантичних образів знайдених анотацій для їх подальшого порівняння та рангування, а також автоматичного поповнення онтології МПС;

онтологія реалізована мовою OWL засобами Protégé API, структура та зміст якої оптимізовані відповідно до інформаційних потреб користувачів;

підсистема класифікації/рангування визначає релевантність до запиту анотацій, доповнених контекстом з онтології на основі розробленого методу оцінювання подібності документів, адаптуючись до результатів класифікації попередніх документів за допомогою регресійного аналізу.

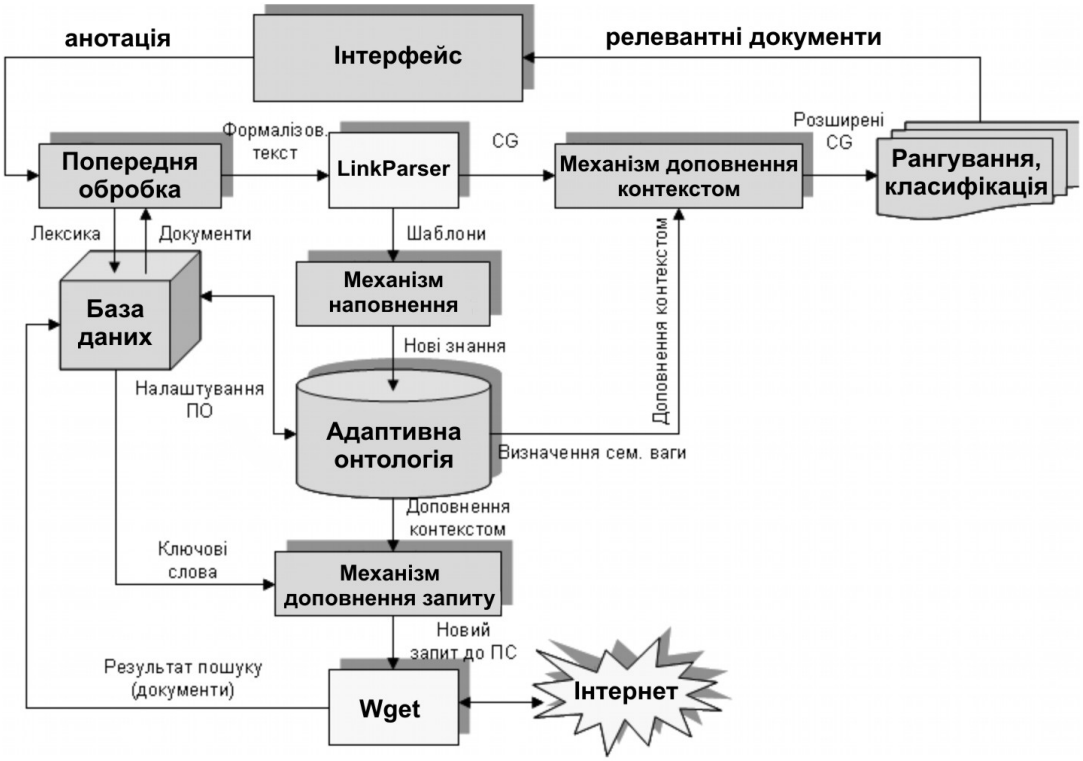


*Рис. 6. Архітектура автоматизованої МПС*

**Висновки**

Побудовано математичну модель адаптації онтології до потреб користувача в складі МПС та розроблено методи оптимізації таких онтологій. Проведено тестування розроблених моделей і проаналізовано отримані результати. З отриманих результатів можна визначити оптимальну кількість елементів у складі онтології. Це дозволило підвищити точність оцінювання подібності текстових документів за їх змістом.

*1. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001. – 383 с. 2. Досин Д. Г., Даревич Р. Р. Побудова базової ланки онтології елемента мультиагентної системи // „Искуственный интелект”. ІПШІ „Наука і освіта”. – 2003. – Вип. 3. – С. 436–444. 3. Даревич Р. Р., Досин Д. Г., В. В. Литвин. Метод автоматичного визначення інформаційної ваги понять в онтології бази знань // Відбір та обробка інформації. – 2005. – Вип. 22(98). – С. 105–111. 4. Dosyn D. G., Darevych R. R., Lytvyn V. V. Modelling of the intelligent text recognition agents based on dynamic ontology. // Тези доп. IV міжнар. конф. „Інтернет – Освіта – Наука – 2004”, Збірник матеріалів конференції. – Вінниця: УНІВЕРСУМ – Вінниця, 2004. – Т. 2. – С. 577–579.*