Методи аналізу наукометричних показників для ранжування дослідників

Керівник:

Доцент кандидат наук Олецький Олексій Віталійович

Виконала студентка КН-4

Горпинюк А.М.

Актуальність

Тема «Методи аналізу наукометричних показників для ранжування дослідників» є актуальною у галузі пошукової оптимізації та ранжування веб-сторінок.

Індекс Гірша – це один з найбільш відомих і популярних методів ранжування науковців, який використовується пошуковими системами для визначення важливості сторінок на основі їх посилань.

Ранжування веб-сторінок є важливим завданням у сфері пошукової оптимізації та інформаційного пошуку. Індекс Гірша та PageRank є одними з найпопулярніших алгоритмів ранжування веб-сторінок. Проте, є багато інших альтернативних підходів, які також можуть бути корисними для розробки нових алгоритмів ранжування та покращення результатів пошуку.

Постановка задачі

- Провести огляд літературних джерел з теорії ранжування веб-сторінок, включаючи індекс Гірша
- Дослідити модуль по роботі з Google Scholar для аналізу алгоритмів ранжування
- Реалізувати консольний застосунок для підрахунку індекса Гірша

Алгоритми ранжування

Індекс Гірша (H-індекс) – це метрика, яка використовується для оцінки науковців на основі кількості їхніх наукових публікацій та кількості цитувань цих публікацій. Він дорівнює n, якщо він має n публікацій, кожна з яких була цитована не менше n разів

PageRank - це алгоритм ранжування веб-сторінок. Альтернативні підходи на основі PageRank використовують цей алгоритм для ранжування наукових статей та авторів замість веб-сторінок.

Індекс CiteRank, який використовує алгоритм PageRank для визначення важливості наукових статей на основі кількості та якості статей, які на них посилаються. Статті з високим індексом CiteRank вважаються важливими, оскільки на них посилаються багато інших статей високої якості.

А також багато інших алгоритмів ранжування()

Робота з даними з Google Scholar

```python
from scholarly import scholarly
list=['L. Page']
for a in list:
    search_query = scholarly.search_author(a)
    author = next(search_query)
    information = (scholarly.fill(author, sections=['basics', 'indices']))
    print('Information about Lyman Page ',information)

```

PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE

```
Information about Lyman Page  {'container_type': 'Author', 'filled': ['indices', 'basics'], 'sopr
urce': <AuthorSource.SEARCH_AUTHOR_SNIPPETS: 'SEARCH_AUTHOR_SNIPPETS'>, 'scholar_id': '_AfgHRoAAAAJ', 'url_picture': 'http
s://scholar.google.com/citations?view_op=medium_photo&user=_AfgHRoAAAAJ', 'name': 'Lyman Page', 'affiliation': 'MIT, Princ
eton University', 'email_domain': '@princeton.edu', 'interests': ['Astronomy', 'Cosmology'], 'citedby': 83127, 'citedby5y'
: 19963, 'hindex': 80, 'hindex5y': 53, 'i10index': 191, 'i10index5y': 137, 'organization': 4836318610601440500, 'homepage'
: 'http://phy-page-imac.princeton.edu/~page/'}
PS D:\python_project>
```

Алгоритм Гірша (C#)

ID	Назва статті	Автор	Рік	Кількість цитувань
1	PageRank: Bringing Order to the Web	L. Page, S. Brin, R. Motwani, T. Winograd	1999	8764
2	The Anatomy of a Large-Scale Hypertextual Web Search Engine	S. Brin, L. Page	1998	5635
3	HITS: A Generalized Framework for Authority Control	J. Kleinberg	1999	2154
4	A Comparison of Document-Level and Passage-Level Link Analysis for Web Search	M. Richardson, P. Domingos	2002	894
5	Link-based Ranking Algorithms	M. Henzinger	2002	546

```
public int CalculateHirschIndex(int authorId)
{
    Author author = Authors.FirstOrDefault(a => a.Id == authorId);
    if (author == null) return 0;

    List<int> citations = author.Papers.Select(p => p.Citations).ToList();
    citations.Sort((a, b) => b.CompareTo(a));

    int hIndex = 0;
    for (int i = 0; i < citations.Count; i++)
    {
        if (citations[i] >= i + 1)
        {
            hIndex = i + 1;
        }
        else
        {
            break;
        }
    }

    return hIndex;
}
```

```
Hirsch index for author (L. Page): 2
Hirsch index for author (S. Brin): 2
Hirsch index for author (R. Motwani): 1
Hirsch index for author (J. Kleinberg): 1
Hirsch index for author (M. Richardson): 1
Hirsch index for author (P. Domingos): 1
Hirsch index for author (M. Henzinger): 1
```

Індекс Гірша має деякі недоліки, включаючи *відсутність* врахування:

- Вищого рівня цитування статей в ядрі Гірша;
- Кількості публікацій, які не увійшли до ядра та рівень їх цитування;
- Інформації про найбільш важливі та високо цитовані роботи;
- Особистого внеску автора (не розрізняються статті з багатьма авторами від статей з одним автором);

Поточні дані з використання *Google Scholar(python)*

```python
from scholarly import scholarly
list=['L. Page', 'S. Brin', 'R. Motwani', 'J. Kleinberg', 'M. Richardson', 'P. Domingos', 'M. Henzinger']
for a in list:
    search_query = scholarly.search_author(a)
    author = next(search_query)
    hirsch = ((scholarly.fill(author, sections=['basics', 'indices']))['hindex'])
    print('Hirsch index for autor',a,': ', hirsch)
```

Результат роботи з модулем Google Scholar

```
Hirsch index for autor L. Page :  80
Hirsch index for autor S. Brin :  90
Hirsch index for autor R. Motwani :  95
Hirsch index for autor J. Kleinberg :  120
Hirsch index for autor M. Richardson :  145
Hirsch index for autor P. Domingos :  87
Hirsch index for autor M. Henzinger :  71
```

Результат консольного застосунку для підрахунку індекса Гірша(C#) на власноруч введеному датасеті

```
Hirsch index for author (L. Page): 2
Hirsch index for author (S. Brin): 2
Hirsch index for author (R. Motwani): 1
Hirsch index for author (J. Kleinberg): 1
Hirsch index for author (M. Richardson): 1
Hirsch index for author (P. Domingos): 1
Hirsch index for author (M. Henzinger): 1
```

Висновки

В результаті виконання кваліфікаційної роботи:

- Написано програмний код який реалізує алгоритм індекса Гірша
- Проведено огляд літературних джерел з алгоритмів ранжування веб-сторінок, включаючи індекс Гірша
- Досліджено, яким чином можна аналізувати дані, які містяться в Google Scholar

Дякую за увагу