УДК 004.44:002.513.5

ЛАНДЕ Д.В., *доктор технічних наук,* керівник наукового центру
НДІ інформатики і права НАПрН України
ЯНЬЦІН ЧЖАО, заступник директора Інституту інформаційних досліджень
Шандуньської академії наук, КНР (ІІД ШАН)
МОЦЗІ ВЕЙ, старший науковий співробітник ІІД ШАН
ШІВЕЙ ЧЖУ, завідувач відділу ІІД ШАН
ЦЗЯНЬПІН ҐО, інженер ІІД ШАН

СИСТЕМА АНОТУВАННЯ КИТАЙСЬКОЇ ПРАВОВОЇ ІНФОРМАЦІЇ

Анотація. *Роботу присвячено методу автоматичного реферування правової інформації КНР, представленої китайською мовою. Розглянуто модель реферату правового документа і принципи його формування. Для вирішення завдання визначення рівня важливості речень з вихідного документа було запропоновано перейти до визначення вагових значень окремих ієрогліфів, а не слів в тексті документів і рефератів.*

Ключові слова: *автоматичне реферування, правова інформація, китайська мова, теорія інформації, мережева модель*

Summary. *Article is devoted to a method of automatic text summarization of the legal information provided in Chinese. The model of the legal document abstract and the procedure of his formation is considered. To determine the level of importance of sentences, it was suggested to proceed to determine the weight values of separate hieroglyphs, rather than words in the text of documents and abstracts. Also consideration of model of documents as networks of sentences for detection of the most important sentences on parameters of this network was offered.*

Keywords: *automatic summarization, legal information, Chinese language, information theory, network model.*

Аннотация. *Работа посвящена методу автоматического реферирования правовой информации КНР, представленной на китайском языке. Рассмотрена модель реферата правового документа и принципы его формирования. Для решения задачи определения уровня важности предложений из исходного документа было предложено перейти к определению весовых значений отдельных иероглифов, а не слов в тексте документов и рефератов.*

Ключевые слова: *автоматическое реферирование, правовая информация, китайский язык, теория информации, сетевая модель*

Постановка проблеми. З постановки завдань штучного перекладу і автоматичного реферування практично починалася комп’ютерна обробка природних мов. Перші фундаментальні роботи з автоматичного реферування текстів з'явилися ще в середині минулого століття [1]. Завдання пов'язане з рішенням найважливішої проблеми – скороченням обсягів інформації, що споживається людиною, боротьби з інформаційним шумом. Це завдання дуже актуальне саме сьогодні через постійне зростання інформаційного простору. Автоматичне реферування відомо всім користувачам мережевих пошукових систем – у відповідь на запит вони отримують не тільки назву документа, але і короткий автоматично створений опис (сніпет), користувачі мобільних пристроїв хочуть бачити короткий опис статей, перш ніж переходять до докладного читання. Особи, які приймають важливі управлінські рішення, повинні ознайомлюватися з тисячами документів на добу, свідомо відкидаючи інформаційний шум.

В даний час існують сотні промислових систем автоматичного реферування, наприклад, такі пакети, як Microsoft Office Word AutoSummarize, Mac OS X Summarize, IBM Tivoli Monitoring Summarization and Pruning Agent, Oracle Text, плагіни для браузерів Chrome, Mozilla.

Відомі численні підходи до автоматичного реферування. Останнім часом, дедалі застосовуються нейромережеві технології, глибинне машинне навчання. Існують також численні лінгвістичні підходи, пов’язані з автоматичним розбором речень, представлених різними мовами. Традиційний тип систем автоматичного реферування – екстрактивний (квазіреферування), при якому реферат складається з окремих, часом слабо пов’язаних між собою речень вихідного документа. Слід зазначити, що сьогодні практично всі промислові системи автоматичного реферування відносяться до екстрактивних систем.

Причин розробки нової системи автоматичного реферування декілька. По-перше, вирішується завдання автоматичного реферування правової інформації. А це тексти, які не можна повною мірою вважати вільними, неструктурованими. Наявна структура окремих видів документів і застосування найкращих універсальних систем реферування не дає задовільних результатів. По-друге, автори мають справу з текстами документів, представленими китайською мовою, що істотно звужує коло можливих для застосування систем. Крім того, для обробки китайських текстів, як правило, необхідна сегментація слів – у китайській мові слова частіше за все не відокремлюються один від одного в тексті. По-третє, має бути розроблена програма, здатна всередині корпоративної системи обробляти великі потоки даних з прийнятною продуктивністю і якістю, вбудована в існуючу систему документообігу.

Крім того, абстрактивний переказ документів в даному випадку неприйнятний. Будь-які “фантазії”, “вольний переказ” комп’ютером правових актів неприпустимі. Вихід виявився один – розробляти деякий гібридний алгоритм і, відповідно, програму екстрактного типу, здатну враховувати особливості правових актів КНР, при цьому програма повинна також бути здатна обробляти окремі документи, які об’єднуються у великі документальні масиви. Ця програма повинна виділяти заздалегідь задані об’єкти в позначених смисловими маркерами частинах документів, виявляти найбільш важливі частини документів (у тому числі і за статистичними критеріями), формувати мережі речень і виводити необхідний обсяг цільової інформації в реферат.

Метою статті є опис нового методу і технології автоматичного реферування правової інформації КНР, представленої китайською мовою.

Виклад основного матеріалу.

Підхід, що пропонується. При вирішенні наведеної проблеми було запропоновано два підходи, які можна вважати новими в даній галузі, а саме, для вирішення завдання визначення рівня важливості окремих речень було запропоновано перейти до визначення вагових значень окремих ієрогліфів, а не слів в тексті документів і рефератів. Також було запропоновано розгляд моделі документів як мережі речень для виявлення найбільш важливих з них за параметрами цієї мережі. Вага зв’язків двох речень у цій мережі визначається нормованою вагою загальних ієрогліфів, що входять в них.

В рамках традиційного статистичного підходу до обробки природних мов вага речень зазвичай обчислюється виходячи з оціночної ваги лексичних одиниць (слів, словосполучень), що входять у ці речення [2 – 5]. В рамках даних робіт пропонується в якості таких елементів для китайської мови використовувати окремі ієрогліфи.

Перехід від розглянутих у класичній моделі слів до ієрогліфів дозволяє уникнути складної процедури сегментування слів у тексті, що неминуче при всіх інших змістовних методах автоматичного аналізу китайських текстів. Звичайно, даний підхід не може бути застосовний до європейських мов, де кількість різних букв не перевищує декількох десятків. Разом з тим для автоматичного реферування китайських текстів запропонований підхід дає прийнятні результати, що буде показано нижче.

Відомо, що в китайській мові існує понад 40 тисяч ієрогліфів, тому кожному з них (нехай окремий ієрогліф не завжди повною мірою відображає смислову одиницю) можна приписати вагове значення, яке розраховується за відомими формулами, наприклад *TF-IDF*, (від англ. *TF* – term frequency, *IDF* – inverse document frequency) [6]. *TF-IDF* – статистична міра, яка використовується для оцінки важливості слова (в даному випадку – не слова, а ієрогліфа) в контексті документа, що є частиною масиву документів. Вага деякого ієрогліфа пропорційна кількості його вживання в документі, і обернено пропорційна частоті появи цього ієрогліфа в усіх документах масиву.

Крім того, на відміну від класичних підходів до визначення вагових значень речень, пропонується нова, мережева модель. В рамках цієї моделі розглядається не спрямована мережа, вузлами якої виступають окремі речення, що входять в документ, між якими встановлюються зв’язки у разі наявності у них загальних ієрогліфів. Вага зв’язку між двома реченнями визначається як сума ваг загальних для цих речень ієрогліфів. На основі цієї мережі розраховується вага кожного речення як сума вагових значень всіх зв’язків, що виходять з відповідного реченню вузла мережі. Природно, вага речень потім нормується, тому що довгі речення без цієї процедури в середньому будуть мати заздалегідь більшу вагу.

Особливості реферування правової інформації. Процедури автоматичного реферування екстрактного класу базуються на визначенні вагових значень (ступенем важливості) окремих речень, які, у свою чергу, залежать від вагових значень слів. У роботі в якості вагових значень слів використовувався класичний критерій *TF-IDF*, хоча це не єдиний можливий для вирішення завдання реферування підхід [7]. Традиційно для визначення вагових значень слів використовувалися два відомих алгоритма – у першому випадку вага речення розглядався як нормована по довжині цього речення сума вагових значень слів, що входять до нього, а у другому випадку використовувався, так званий, алгоритм симетричного реферування [8]. У цьому випадку вага речення визначається як сума вагових значень його зв’язків з попереднім і наступним реченнями.

Крім того, в даній роботі запропоновано мережевий алгоритм, в якому на відміну від другого випадку обчислюються зв’язки не тільки між сусідніми реченнями, а й між усіма реченнями у тексті документа. Такий підхід, звичайно, обчислювально більш складний, ніж перші два, однак, як показала практика, призводить до кращих результатів. При цьому складність алгоритму, в разі розглянутого підходу реферування текстів, наведених китайською мовою, компенсується тим, що замість слів (сегментація яких в даному випадку не потрібна) розглядаються лише окремі ієрогліфи.

Слід зазначити, що специфіка правової інформації, вимоги до структури і обсягу реферату, дозволили використовувати наведені вище універсальні підходи до вирішення окремої спеціальної задачі.

До структури і обсягу реферату правового документа (приклади таких документів можна знайти на сайті http://www.gov.cn в розділі /zhengce) висуваються вимоги, які знайшли свою програмну реалізацію:

1. Реферат починатися з заголовка документа, наведеного практично без змін.

2. У рефераті відзначається вид документа (оголошення “通告”, звіт “报告”, результати роботи “工作 成果”, положення “政策” тощо).

3. Якщо у документі позначена його мета (маркери: “目的”, “奖补目的”, “调整目的”, “普查的目的和意义” тощо), то вона також знаходить відображення у рефераті.

4. Якщо в першому або другому реченні документа позначені суб’єкти призначення документів (що також видно за спеціальними маркерами), то таке речення також включається до складу реферату.

5. Якщо в заголовку документа або в позначенні його мети в явному вигляді присутні об’єкти з числа заздалегідь відомих (що входять у таблицю базових об’єктів), то ці об'єкти повинні бути виділені в рефераті.

6. Якщо документ відноситься до типу, що не підлягає подальшому реферуванню (нагороди “表彰”, оголошення про торги “招标”, листи “函” і ін), то реферат також вважається підготовленим.

7. З тексту документа вибираються всі речення, що містять вибрані із заголовка і цілі об’єкти. Якщо таких речень менше необхідного числа (яке задається заздалегідь або розраховується виходячи з обсягу документа), то вони виводяться в реферати в тій же послідовності, що і в первинному документі. Реферат вважається підготовленим.

8. Якщо речень більше необхідного числа, то вони зважуються відповідно до наведеного вище алгоритму (за результатами тестування обрано мережевий алгоритм). Після цього речення ранжируються за вагою і необхідна їх кількість виводиться в реферат в тій же послідовності, що і в первинному документі. Реферат вважається підготовленим.

Згідно з наведеними вимогами була розроблена програма автоматичного реферування правової інформації, наданої китайською мовою.

Суміжні завдання Text Mining. Автоматичне реферування текстів – це одна з важливих задач технологій глибинного аналізу текстів (Text Mining), яка включає ще декілька напрямків, таких як витяг сутностей (Information extraction), побудова мереж слів (Language Networks), що відображають особливості предметних областей, кластеризації (Cluster Analysis).

Запропонований для реферування алгоритм спирається на деяку множину заздалегідь підготовлених слів, що відображають основні об’єкти, представлені в правових документах (наприклад, “人口” – населення, “产业” – промисловість, “儿童” – діти, тощо).

Разом з тим, якщо застосувати алгоритм сегментації слів, після чого їх ранжирувати, то легко можна виділити “розширення” стартових об’єктів, що найбільш часто зустрічаються, наприклад, поняття “організація” (组织) розширити до поняття “міжнародна організація” (国际 组织), “громадська організація” (社会 组织), а поняття “оборона” (事业) до поняття “народна протиповітряна оборона” (人民 防空 事业). У результаті документам масиву правової інформації були поставлені у відповідність основні поняття, які можуть виступати в якості “ключових слів”, дескрипторів, основ побудови моделей предметних областей (Subject Domain).

Як один з видів моделей предметних областей може розглядатися мережа слів, вузли якої відповідають окремим поняттям [9]. Були запропоновані і реалізовані такі прості правила побудови цієї мережі, тобто правила встановлення зв’язків між вузлами:

1. Всі об’єкти з базового, заздалегідь підготовленого списку, що входять в один документ зв’язуються зв’язками.
2. Якщо два об’єкти входять до N різних документів, то сила зв’язку між ними дорівнює N.
3. Поняття, що є розширеннями понять з стартового набору, зв’язуються з відповідними базовими поняттями.

За допомогою програми Gephi (http://gephi.org) [10] побудована мережа була візуалізована (Рис. 1) і були отримані такі парметри побудованої мережі: кількість вузлів – 3364 (кількість об’єктів з стартового набору – 220); кількість зв’язків – 10167; щільність мережі – 0.001; кількість зв’язаних компонент – 6; середня довжина шляху – 3.013; середній коефіцієнт кластеризації – 0.859.

До топологічної особливості побудованої мережі відноситься дуже великий середній коефіцієнт кластеризації. Це пояснюється, з одого боку, великою кількістю понять, пов’язаних лише з поняттям, що їх породжує (відсутність інших сусідів), а з іншого боку сильною зв’язністю об’єктів зі стартового списку. Невелика середня довжина шляху свідчить про те, що ця мережа є “малим світом” (Small World) [11].

Наведене на Рис. 1 та 2 загальний вигляд мережі слів наочно демонструє подальшу можливість кластеризації мережі, вибору підмножин – кластерів із слів (понять). Ця процедура дозволяє виділяти тематичні підмножини в рамках розглянутої предметної області.

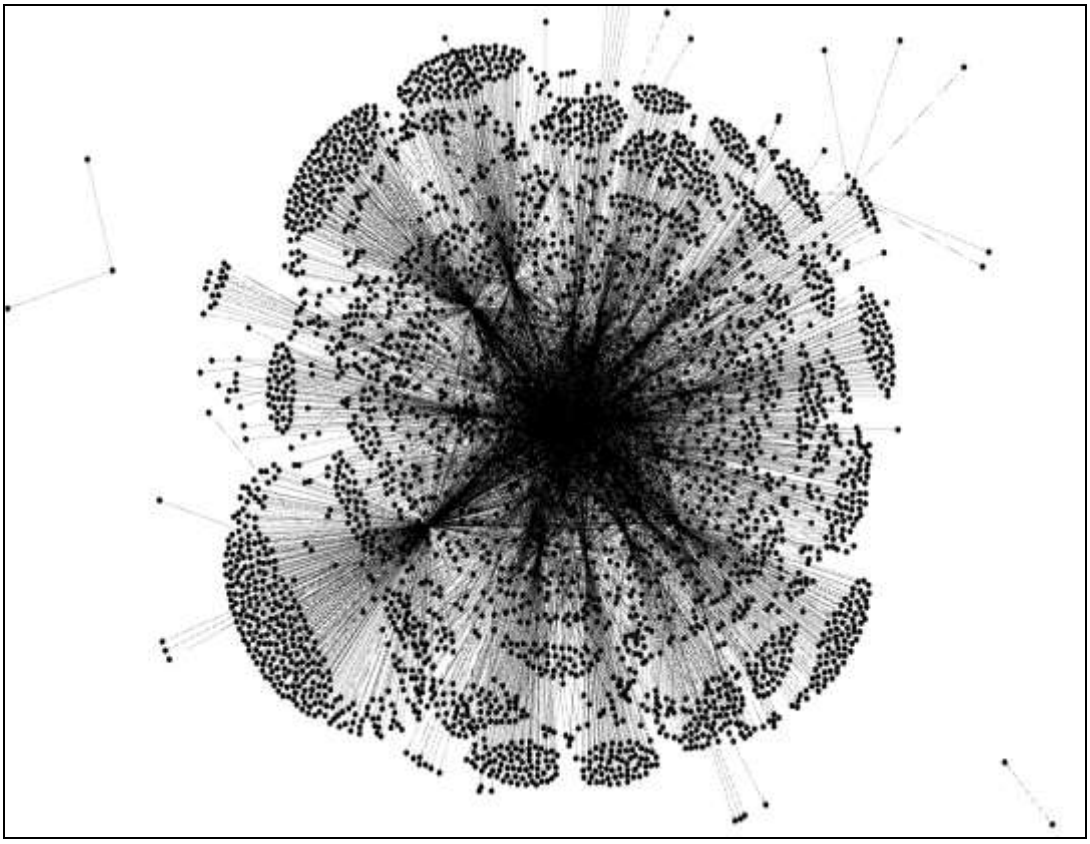
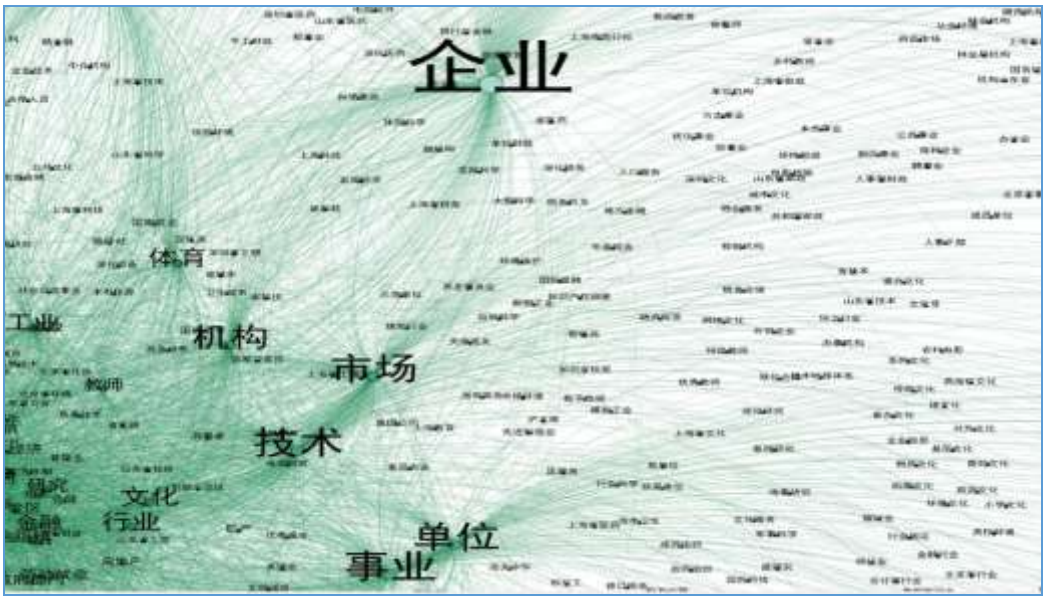


Рис. 1 – Загальний вигляд мережі слів



Фрагмент мережі слів

Рис. 2 – Мережа слів, що відображує предметну область

За допомогою програми Gephi також були отримані списки найбільш вагомих вузлів відповідно до мережевих рангових критеріїв PageRank і HITS [12] (Рис. 3).

Label	PageRank
水利	0.001548
“十一五”	0.00154
扶贫	0.00149
毕业生	0.001408
银行	0.001405
上海市财政	0.001383
邮政	0.001374
林业	0.001352
信息传输	0.001349
城市规划	0.001337
文物	0.001323
医生	0.001315
省财政	0.001242
技术服务	0.001215
山东省财政	0.001194
农民工	0.001126
县域	0.00111
农田	0.001097
电信	0.001072
经济特区	0.001055
科技创新中心	0.001045
试验区	0.001
北京市财政	0.000989
食品药品	0.000964
电影	0.000949
房地产	0.000897
矿产	0.00088
供销	0.000877

PageRank

Label	Hub
残疾人	0.04637
租赁	0.044748
科技创新中心	0.043809
农民工	0.043738
统计	0.04244
电信	0.04112
省财政	0.040688
经济特区	0.039118
山东省财政	0.039081
作业	0.038172
食品药品	0.036646
北京市财政	0.036476
水利	0.036002
试验区	0.035958
电影	0.031812
人工智能	0.031356
娱乐	0.03121
邮政	0.028793
物流业	0.027323
海关	0.026766
社会信用体系建设	0.026739
餐饮	0.02619
深圳市市场	0.02569
干部	0.025645
公共管理	0.025336
金融业	0.025252
食品药品监管	0.025083
经济体制	0.025021

HITS

Рис. 3 – Найбільш рейтингові слова за критеріями PageRank і HITS

Оцінювання результатів. Для оцінювання результатів застосовується дві оцінки якості реферату без участі експертів – косинусна міра і дивергенція Дженсена-Шеннона (Jensen-Shannon), обґрунтування застосування яких надано в роботі [13].

Строго кажучи, міра Дженсена-Шеннона відповідає втраті інформації при реферування і пропорційна сумарній вазі слів (в нашому випадку – ієрогліфів), що входять в документ, але відсутні в рефераті.

При реферування була реалізована нова ідея визначення вагових значень речень на основі вагових значень окремих ієрогліфів, а не слів, як це загальноприйняте. Тому якість реферування перевірялася не лише виходячи з урахування ваги окремих ієрогліфів, а й з урахуванням ваги цілих слів, що входять в документи і реферати, щоб переконатися, що запропонований підхід задовільний і за критеріями традиційних систем реферування. Природно, для цього довелося виконати витратну за ресурсами процедуру сегментації слів [14]. Слід зазначити, що дана процедура виконувалася виключно для перевірки якості алгоритмів реферування і не входить до складу самого алгоритму реферування.

Випробування проводилися на реальному масиві правової інформації Китайської народної республіки обсягом 10 тисяч документів.

Висновки.

Результати випробувань дозволяють резюмувати наступне:

1. В роботі представлена гібридна методика автоматичного реферування, що охоплює статистичні та маркерні методи, а також облік розташування речень у тексті правового документа. Запропонована модель реферату відображає інформаційну потребу замовників при роботі з правовою інформацією. Наведені підходи призводять до результатів, якість відповідає представленим на відомій конференції з аналізу текстів.

2. Реалізовано підхід до визначення вагових значень окремих ієрогліфів, а не сегментованих слів в тексті документів. Дана методика дозволяє уникати витратної процедури сегментування слів, необхідної для інших змістовних методів обробки текстів, наведених китайською мовою.

3. Реалізовано і випробувані різні методи автоматичного реферування. Реферування на основі мережевої моделі документа виявилося кращим за критеріями косинусної міри і відстані Дженсена-Шеннона для рефератів, обсяг яких перевищує 2 речення.

4. Запропонований підхід з урахуванням змін в маркерах-шаблонах може використовуватися не тільки для правових документів, а й для текстів довільної тематики, зокрема, науково-технічної та новинної інформації.

Використана література

1. Luhn Hans Peter. The automatic creation of literature abstracts // IBM Journal of research and development. – 1958. – № 2. – Pp. 159-165.

2. Zhang C. Automatic Keyword Extraction from Documents using Conditional Random Fields // Journal of Computational Information Systems. – 2008. – № 4 (3). – Pp. 1169-1180.

3. Ramos J. Using tf-idf to determine word relevance in document queries / Proceedings of the first instructional conference on machine learning, 2003. – Pp. 1-4.

4. Bhart, Santosh Kumar, Babu Korra Sathya, Pradhan, Anima. Automatic Keyword Extraction for Text Summarization in Multi-document e-Newspapers Articles // European Journal of Advances in Engineering and Technology. – 2017. – 4 (6). – Pp. 410-427.

5. Chien L.-F. Pat-tree-based keyword extraction for Chinese information retrieval / ACM SIGIR Forum. 31, ACM, 1997. – Pp. 50-58.

6. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval / Information Processing & Management. – 1998. – 24(5). – Pp. 513-523.

7. Lande D.V., Snarskii A. A, Yagunova E.V., Pronoza E. V. The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text / 12th Mexican International Conference on Artificial Intelligence, 2013. – Pp. 209-215. DOI: 10.1109/MICAI. 2013.33

8. Яцко В.А. Симметричное реферирование : теоретические основы и методика // Научно-техническая информация. – (Серія 2). – 2002. – № 5. – С. 18-28.

9. Ланде Д.В. Елементи комп’ютерної лінгвістики в правовій інформатиці. – К. : НДІІП НАПрН України, 2014. – 168 с. ISBN 978-966-2344-33-2

10. Cherven Ken. Network Graph Analysis and Visualization with Gephi. – Packt Publishing, 2013. ISBN: 9781783280131

11. Kleinberg J. Navigation in a small world // Nature. – 2000. – № 406 (6798). – Pp. 845. DOI: 10.1038/35022643

12. Langville Amy N., Meyer Carl D. Google’s PageRank and beyond: the science of search engine rankings. – Princeton university press, 2011. ISBN: 9780691152660

13. Louis Annie, Nenkova Ani. Automatic Summary Evaluation without Human Models / In First Text Analysis Conference (TAC’08). – Gaithersburg, MD, Etats-Unis, 17-19 November 2008.

14. Berezin Boris A., Lande Dmitry V., Pavlenko Oleh Y. Development, Evaluation and Usage of Word Segmentation Algorithm for National Internet Resources Monitoring Systems / CEUR Workshop Proceedings, 2017. Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017). – 2067. – Pp. 16-22.

~~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~~
~~~~~~~~~~~~~~