Міжнародна науково-технічна конференція

ІНТЕЛЕКТУАЛЬНІ ТЕХНОЛОГІЇ ЛІНГВІСТИЧНОГО АНАЛІЗУ

23-24 жовтня 2013 року

ПІДХОДИ ДО АВТОМАТИЧНОГО ВИЗНАЧЕННЯ ТЕРМІНОЛОГІЧНИХ ОСНОВ ОНТОЛОГІЙ

Д.В. ЛАНДЕ, д.т.н.,
завідувач відділом
Інституту проблем реєстрації інформації НАН України

Київ, 23 жовтня 2013 р.

Анотація

Доповідь присвячено актуальному на цей час питанню визначення важливих структурних елементів тексту, що виявляються інформаційно-значущими, визначають інформаційну структуру. Використання таких елементів дозволяє формувати тезауруси, пошукові образи документів, онтології.

Існуючі підходи

Ключові слова для пошуку в тексті, опорні слова для автоматичного екстрагування значущих фрагментів текстів або формування автоматичних рефератів, вибираються з урахуванням такої властивості слів, як «дискримінантна сила», для визначення яких існує три класичних методи – TFIDF, дисперсійний, ентропійний. Також застосовуються й новітні мережеві методи.

Показник TFIDF

Для кожного терміну з текстового корпусу у канонічному виді TFIDF дорівнює добутку частоти слова у фрагменті тексту (Term Frequency) на логарифм від величини, зворотної кількості окремих фрагментів тексту, в яких це слово зустрілось (Inverse Document Frequency). СЕРЕДНЄ ЗА МАСИВОМ. Для кожного слова i, що входить до текстового корпусу, який складається з N документів, підраховується кількість документів df(i), в яких міститься це слово, а також загальна частота входження даного слова у текстовий корпус – n(i). Після цього розраховується середнє значення TFIDF для кожного слова:

$$tfidf(i) = \frac{n(i)}{N} \log\left(\frac{N}{df(i)}\right)$$

TFIDF Okapi BM25

TFIDF дає не зовсім коректні результати на текстових масивах із документів з різною за довжиною. Тому запропоновано його модифікацію – Okapi BM25:

$$TF \times IDF(w_i) = IDF(w_i) \frac{f(w_i, D) \times (k_1 + 1)}{f(w_i, D) + k_1 \times (1 - b + b \times \frac{|D|}{avgdl})}$$

де f(wi,D) – частота слова у документі D; |D| – довжина документа D; avgdl – середня довжина документа в колекції; k1 і b – параметри, що вибираються експертно. IDF обчислюється за формулою:

$$IDF(w_i) = \log \frac{N - n(w_i) + 0.5}{n(w_i) + 0.5},$$

де N – загальна кількість документів у масиві, n(wi) – кількість документів, що містять термін wi.

Дисперсійна оцінка

Дисперсійна оцінка важливості термінів обчислюється для окремих термінів наступним чином. Позначимо середній інтервал між появами слова A у тексті через $\langle \Delta A \rangle$, а середній квадрат значень цих інтервалів через $\langle \Delta A^2 \rangle$. Дисперсійна оцінка слова σ_A розраховується як

$$\sigma_A = \frac{\sqrt{\langle \Delta A^2 \rangle - \langle \Delta A \rangle^2}}{\langle \Delta A \rangle}.$$

Ідея дисперсійної оцінки близька до TFIDF, однак більш коректно застосовується до цілих текстів, а не до масивів з великої кількості документів, як TFIDF.

Оцінка PMI

Якщо розглядати терміни як події, можемо знайти так званну ентропійну оцінку ваги терміну a (Pointwise Mutual Information, PMI):

$$m(w) = -\log_2 \frac{p(w \mid d)}{p(w)} = -\log_2 \frac{n_{w,d}}{n_w},$$

де $n_{w,d}$ – кількість входжень терміну до тексту документа, n_w – загальна кількість входжень.

Мережі мови

Першим кроком при застосуванні теорії складних мереж до аналізу тексту є представлення цього тексту у вигляді сукупності вузлів і зв'язків, побудова мережі мови (Language Network).

Існують різні інтерпретації значень вузлів і зв'язків, що веде, відповідно до різних представлень мережі мови.

Вузли можуть бути поєднані між собою, якщо відповідні їм слова знаходяться поруч у тексті, належать одному реченню, поєднані синтаксично або семантично.

Простійші мережі мови

L-простір. Зв'язуються сусідні слова, що належать одному реченню. Кількість сусідів для кожного слова (вікно слова) визначається радіусом взаємодії R, найчастіше розглядається випадок R = 1.

B-простір. Розглядаються вузли двох типів, що відповідають реченням і словам, які входять до них.

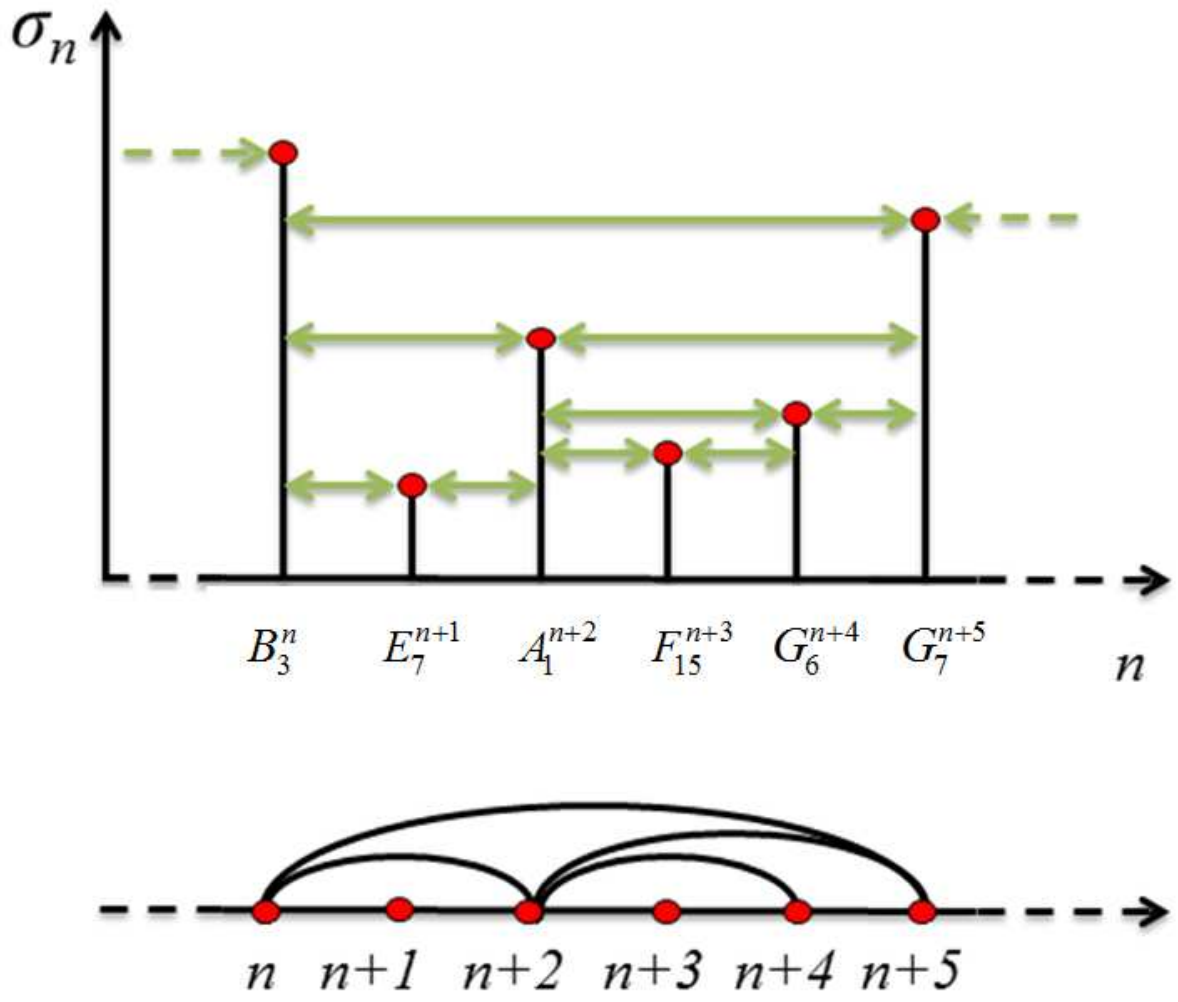
P-простір. Всі слова, що належать одному реченню, зв'язуються між собою.

C-простір. Речення зв'язуються між собою, якщо у них застосовуються однакові слова.

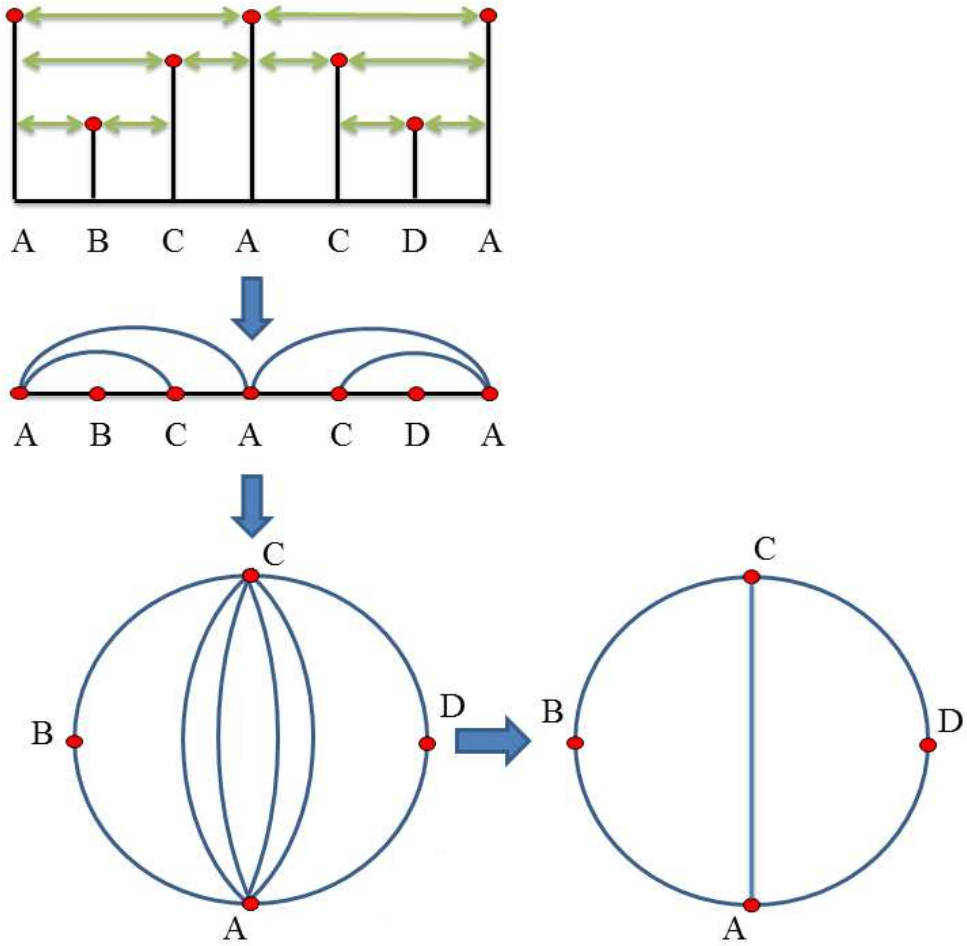
Побудова графів на основі рядів значень

На цей час запропоновано декілька методів побудови мереж на основі часових рядів, серед яких можна назвати методи побудови графів видимости, зокрема так званний граф горизонтальної видимости (Horizontal Visibility Graph – HVG). Ці подходи дозволяють будувати мережеві структури також і на основі текстів, у яких окремим словам або словосполученням деяким чином поставлені у відповідність деякі вагові значення.

Формування графа горизонтальної видимости



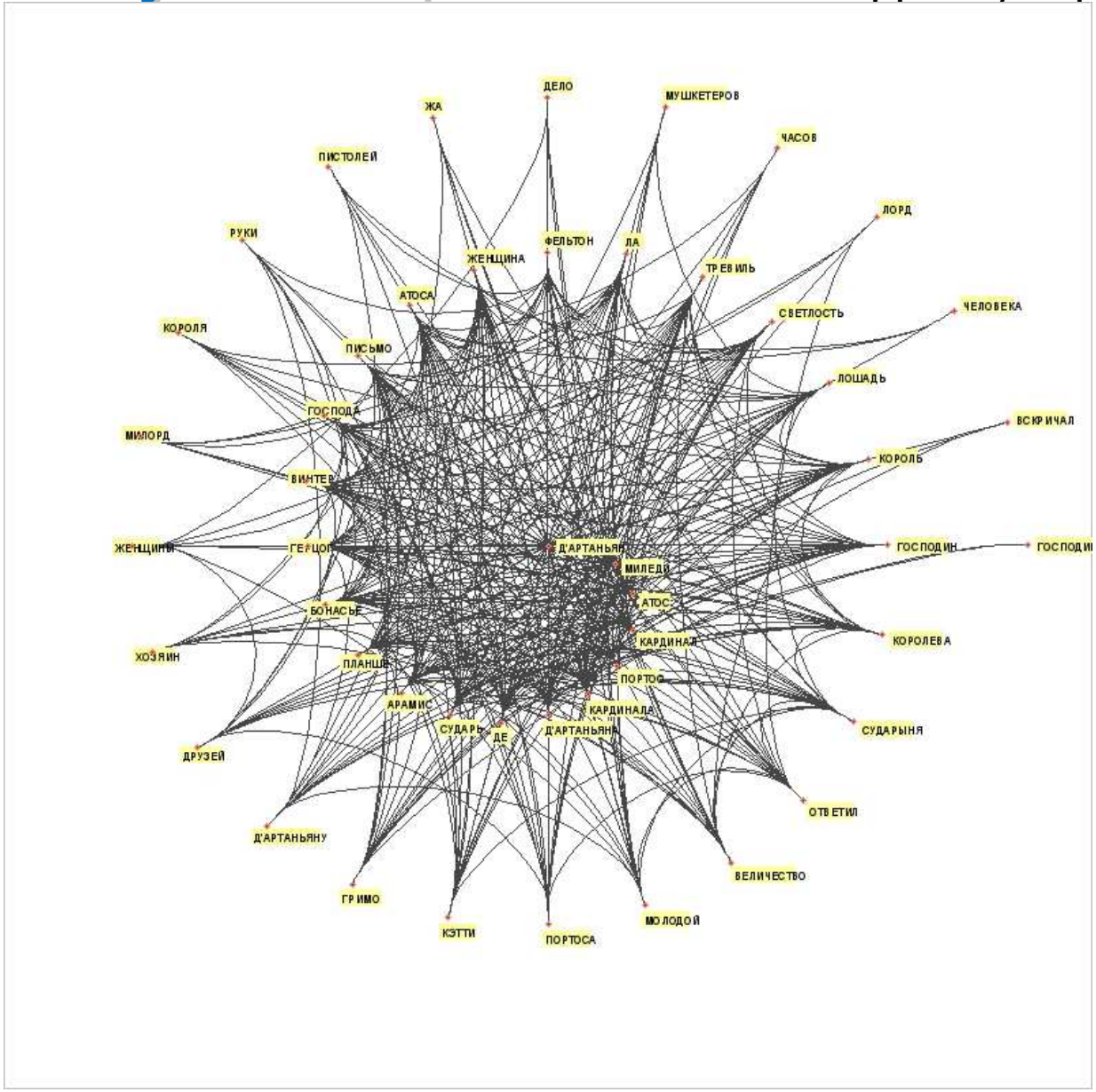
Етапи побудови КГГВ



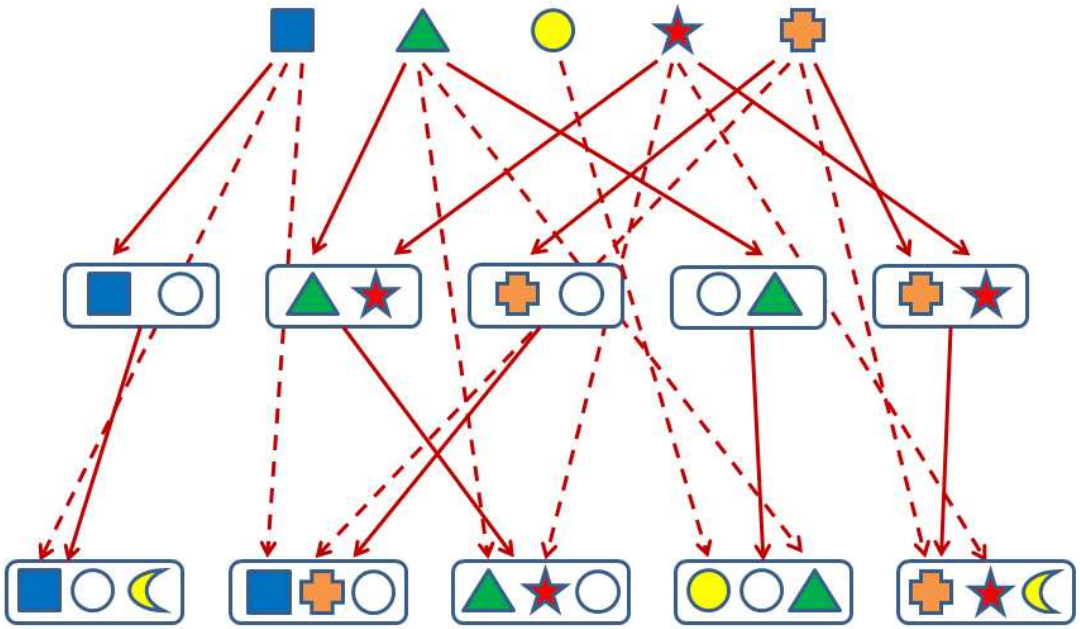
Компактифікований граф горизонтальної видимості

Візуалізація КГГВ

О. Дюма, «Три мушкетери»



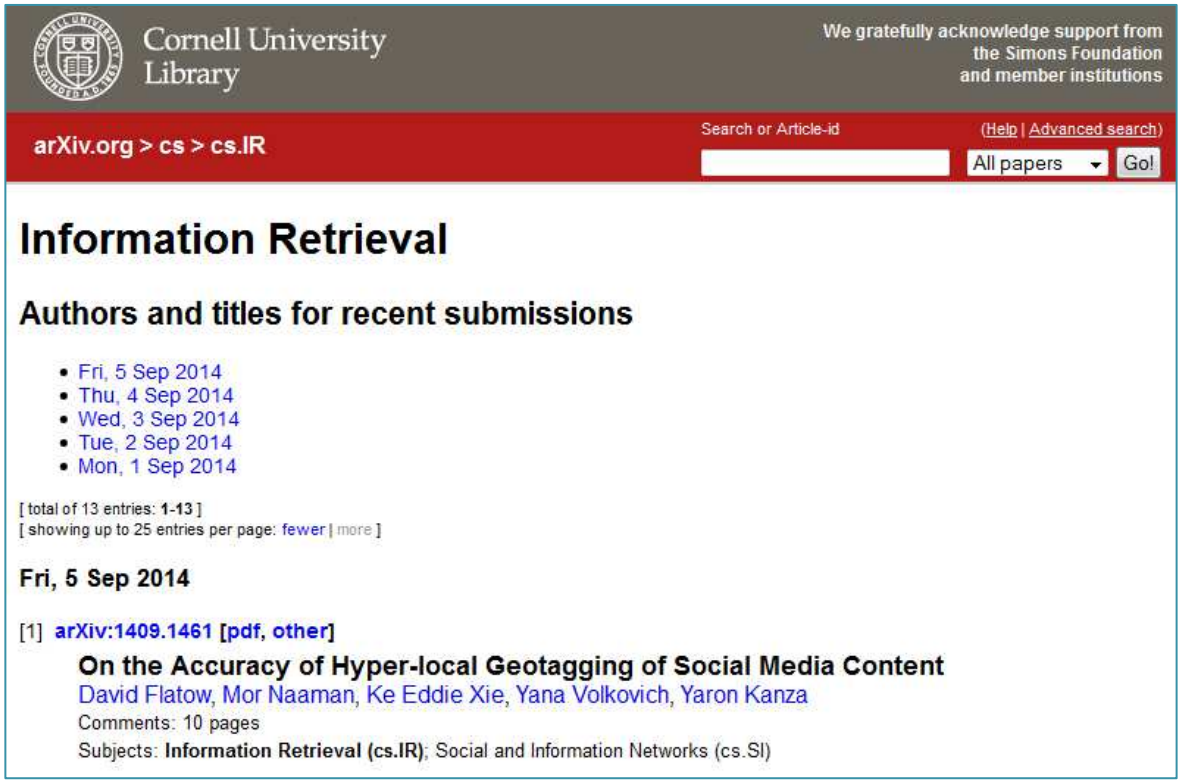
Мережі природних ієрархій термінів



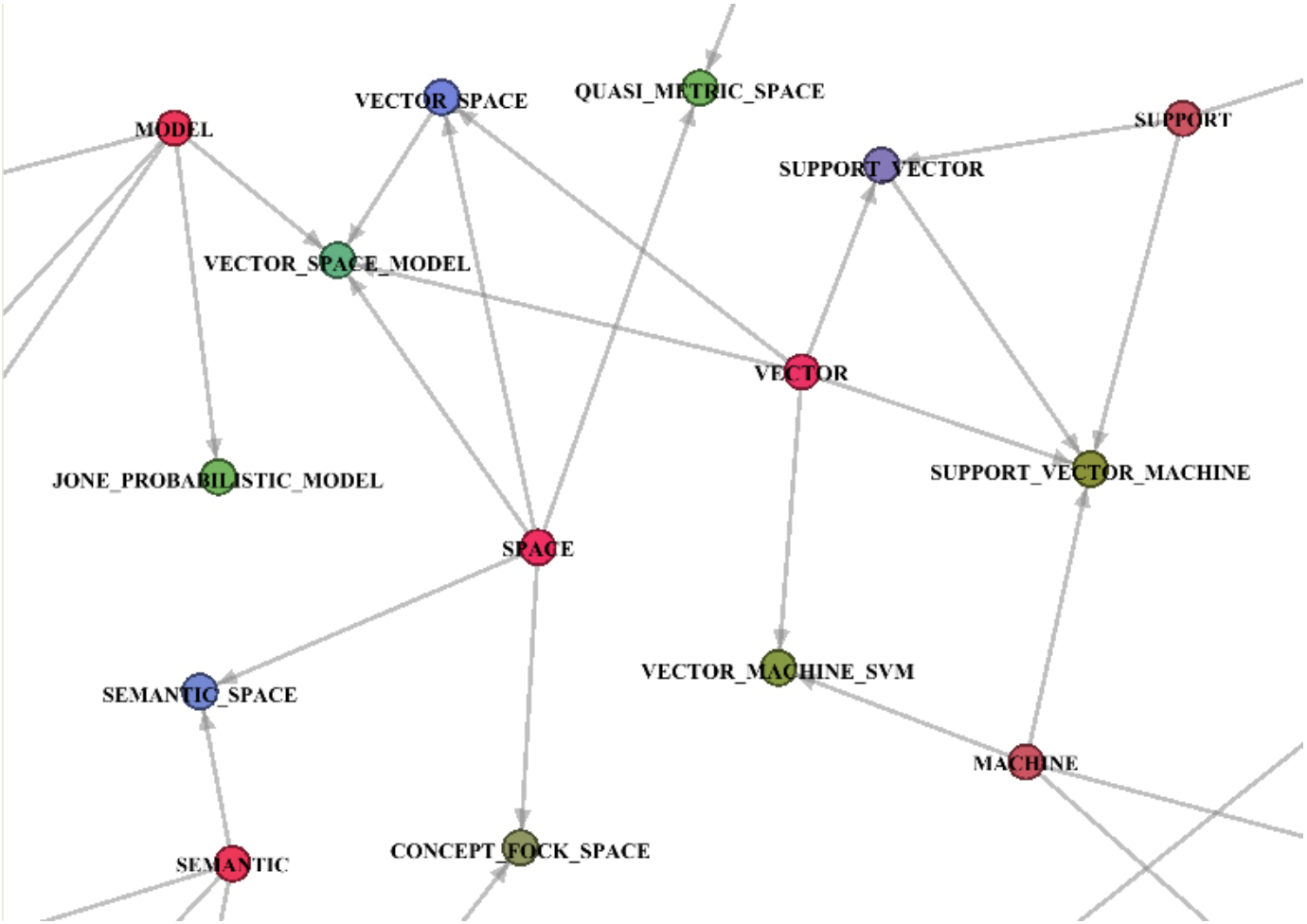
З відібраних термінів (уніграм, біграм, триграм) формуються мережі природних ієрархій термінів, у яких як вузли розглядаються самі терміни, а зв'язкам відповідають входження одних термінів у інші.

Вихідний текстовий корпус

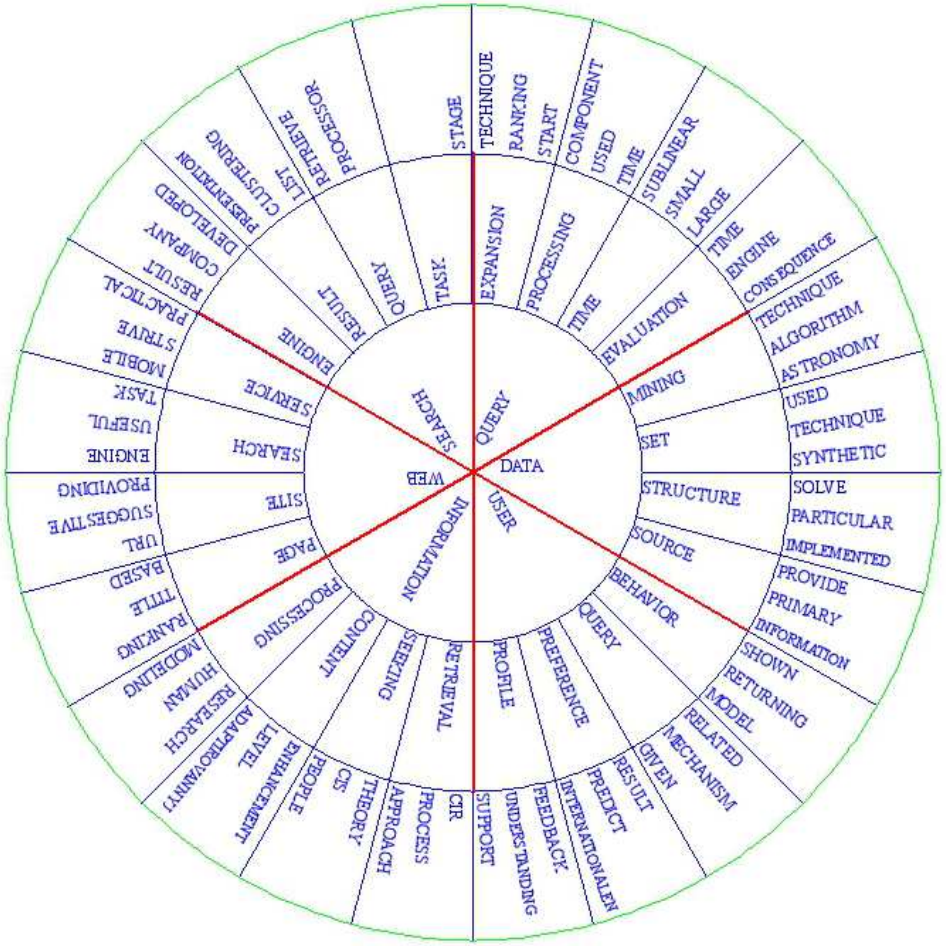
Як приклад вихідного текстового корпусу розглядаються анотації електронних препринтів Arxiv (http://arxiv.org) за тематикою інформаційного пошуку (рубрика cs.IR, понад 500 документів за 5 років).

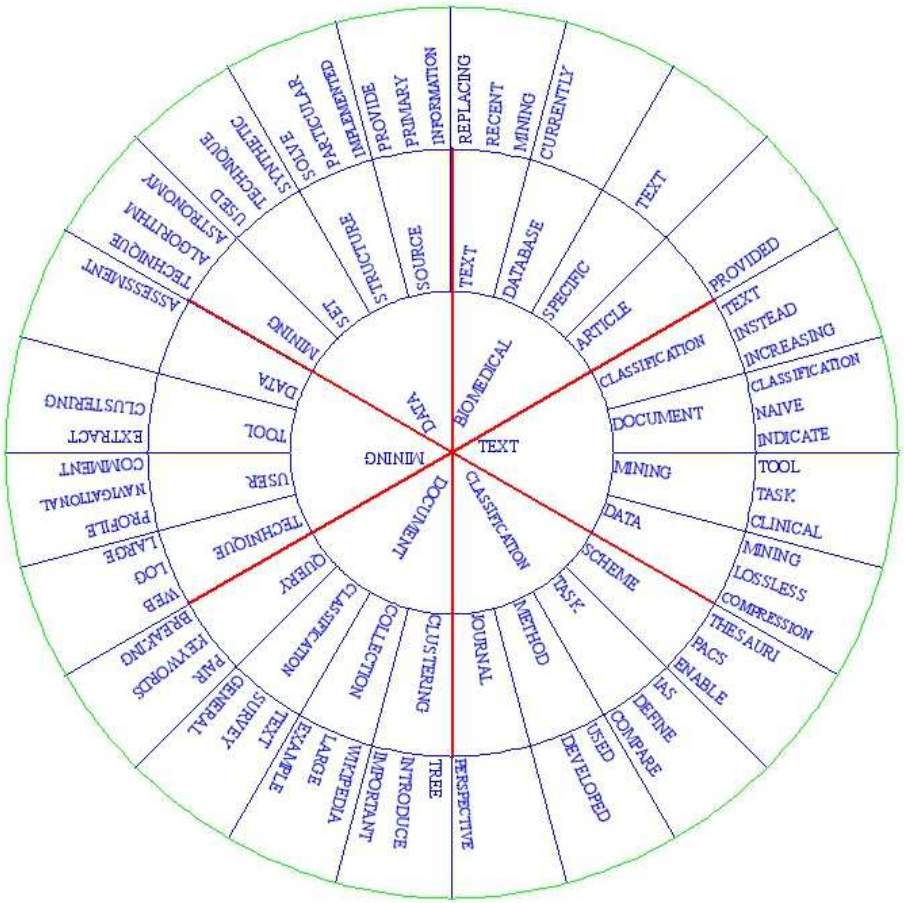


Візуалізація фрагменту МПІТ



Радиальні (секторні) діаграми





Ранжирування вузлів МПІТ

За допомогою алгоритму HITS забезпечується вибір з мережі кращих «авторів» (вузлів, на які ведуть посилання) і «посредників» (вузлів, від яких йдуть посилання включення). Термін є добрим посередником, якщо від нього йдуть зв'язки на важливі словосполучення, і навпаки, добрим автором, якщо на нього ведуть зв'язки від важливих авторів. У відповідності з алгоритмом HITS для кожного вузла мережі vj рекурсивно обчислюється його значимість як автора a(vj) і посередника h(vj) за формулами:

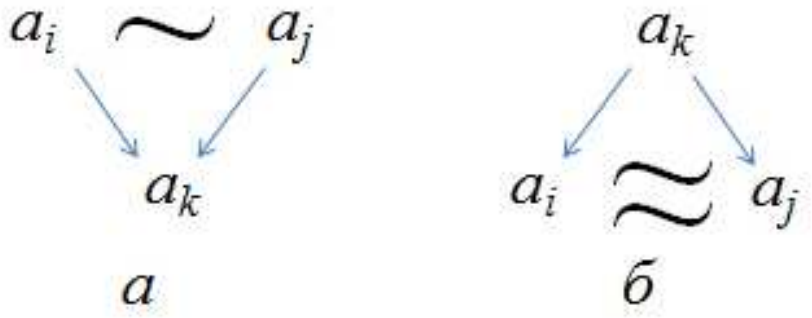
$$a(v_j) = \sum_i h(v_i);\ h(v_j) = \sum_i a(v_i).$$

Терміни з найкращим авторством/посередництвом

Авторы	Посредники
WEB_SEARCH_ENGINE	SEARCH
SEARCH_ENGINE_RESULT	ENGINE
BASED_SEARCH_ENGINE	SEARCH_ENGINE
VERTICAL_SEARCH_ENGINE	WEB
SEARCH_ENGINE_COMPANIE	RESULT
GEOGRAPHIC_SEARCH_ENGINE	SEARCH_RESULT
ENTITY_SEARCH_ENGINE	WEB_SEARCH
SEARCH_RESULT_CLUSTERING	IMAGE
TAG_SEARCH_RESUL	PAGE
SEARCH_RESULT_PRESENTATION	CLUSTERING
PERSONALIZE_SEARCH_RESULT	TAG

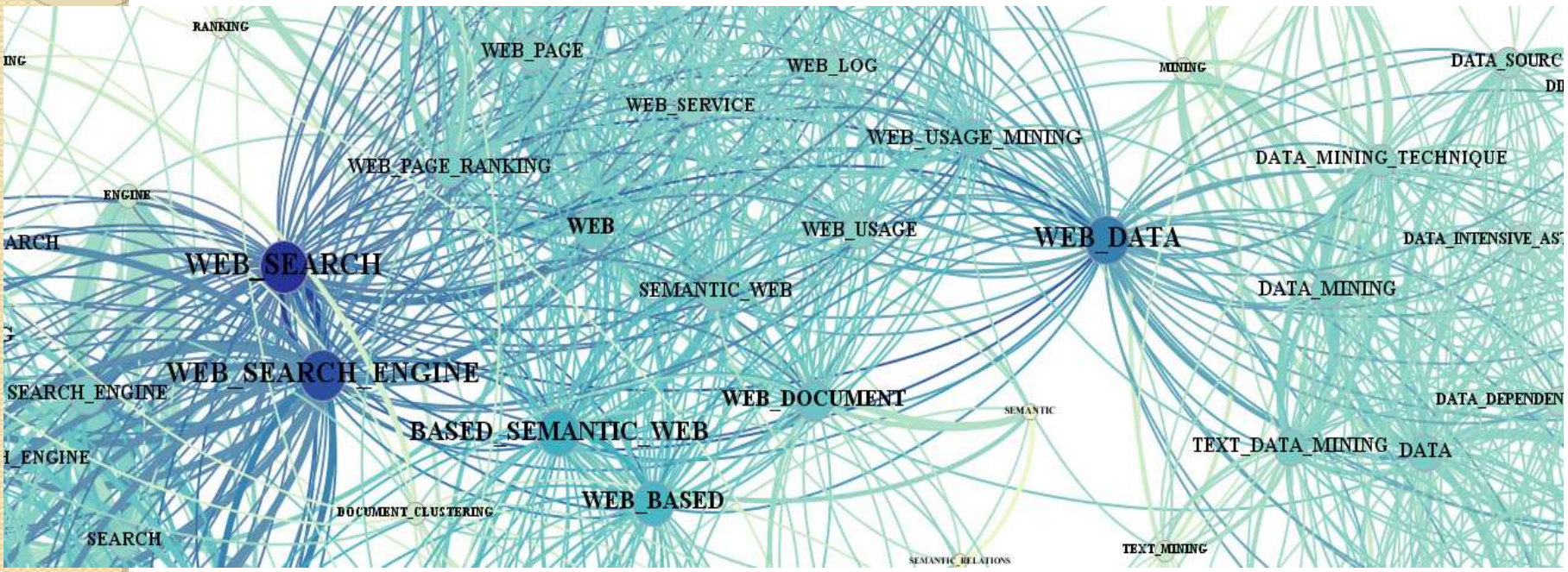
Асоціативні зв'язки

Мережу МПІТ СЕИТ можна розглядати як основу для формування зв'язків між її вузлами. Якщо позначити матрицю інцидентності МПІТ літерою A, то матриці AA^T і A^TA будуть відображувати зв'язки таких типів:



а *б*

Фрагмент МПІТ з асоціативними зв'язками



Деякі висновки

- Зроблено огляд існуючих алгоритмів виявлення ключових слів.
- Запропоновано алгоритм побудови мереж природних ієрархій термінів.
- Запропоновано методи візуалізації фрагментів МПІТ.
- Запропоновано алгоритм побудови асоціативних зв'язків між термінами в МПІТ.
- Запропоновано використання алгоритму HITS для вибору найбільш важливих елементів в МПІТ.
- Мережу мови, що побудовано за допомогою запропонованої методики, можна використовувати як базу для побудови онтологій предметної області.

Міжнародна науково-технічна конференція

ІНТЕЛЕКТУАЛЬНІ ТЕХНОЛОГІЇ ЛІНГВІСТИЧНОГО АНАЛІЗУ

23-24 жовтня 2013 року

Дякую за увагу!

Д.В. ЛАНДЕ

dwlande@gmail.com