Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

Identyfikacja wybranych typów podstruktur w sieciach reprezentujących opinie

Michał Ochman, WFiIS AGH
opiekun: dr hab. Jarosław Kwapień, IFJ PAN

Kraków, 27 września 2013

Plan prezentacji

Cel pracy

Eksploracja danych

Przetwarzanie języka naturalnego

Sieci złożone

Małe światy

Rozkład stopni wierzchołków

Upraszczanie sieci

Szacowanie opinii

Collaborative similarity

Motywy

Cel pracy

Przedmiotem zainteresowania pracy jest struktura i ewolucja sieci opisujących opinie wybranych grup agentów.

Cel pracy

Przedmiotem zainteresowania pracy jest struktura i ewolucja sieci opisujących opinie wybranych grup agentów.

Źródła problemów:

- Marketing szeptany (word of mouth).
- Promowanie mody (trend setting).
- Marketing społecznościowy (community marketing).

Cel pracy

Przedmiotem zainteresowania pracy jest struktura i ewolucja sieci opisujących opinie wybranych grup agentów.

Źródła problemów:

- Marketing szeptany (word of mouth).
- Promowanie mody (trend setting).
- Marketing społecznościowy (community marketing).

Ze względu na zaufanie „zwykłych” użytkowników do „guru” istnieje potrzeba wyławiania nieszczerych użytkowników.

Eksploracja danych

Zwykle odkrywanie wiedzy z istniejących baz danych, ale nie tylko.
Tu, przetwarzanie jednego typu zbioru danych w inny.

Eksploracja danych

Zwykle odkrywanie wiedzy z istniejących baz danych, ale nie tylko.
Tu, przetwarzanie jednego typu zbioru danych w inny.

Analiza danych:

- Asocjacje.
- Klasteryzacja.

Przetwarzanie języka naturalnego

Segmentacja i normalizacja

Ach, to on! $\rightarrow$ [ach; to; on]

Przetwarzanie języka naturalnego

Segmentacja i normalizacja

Ach, to on! $\rightarrow$ [ach; to; on]

Klasyfikacja

- $\times$ *Stopwords*.
- $\times$ Funktory.
- ✓ Nazwy.

Przetwarzanie języka naturalnego

Segmentacja i normalizacja

Ach, to on! $\rightarrow$ [ach; to; on]

Klasyfikacja

- $\times$ *Stopwords*.
- $\times$ Funktory.
- ✓ Nazwy.

Szukanie rdzenia (lub podstawy słowotwórczej)

- Usuwanie przyrostków

Sieci złożone

Grafy:

- o nieregularnych i nielosowych połączeniach,
- o nietrywialnej topologii:
  - charakterystyczny rozkład stopni wierzchołków,
  - wysoki współczynnik klastrowania,
  - asortatywność.

Sieci złożone

Grafy:

- o nieregularnych i nielosowych połączeniach,
- o nietrywialnej topologii:
  - charakterystyczny rozkład stopni wierzchołków,
  - wysoki współczynnik klastrowania,
  - asortatywność.

Typy:

- Sieci społecznościowe
- Sieci informacyjne
- Sieci biologiczne

Małe światy

Eksperyment Milgrama – sześć stopni oddalenia.

Małe światy

Eksperyment Milgrama – sześć stopni oddalenia.

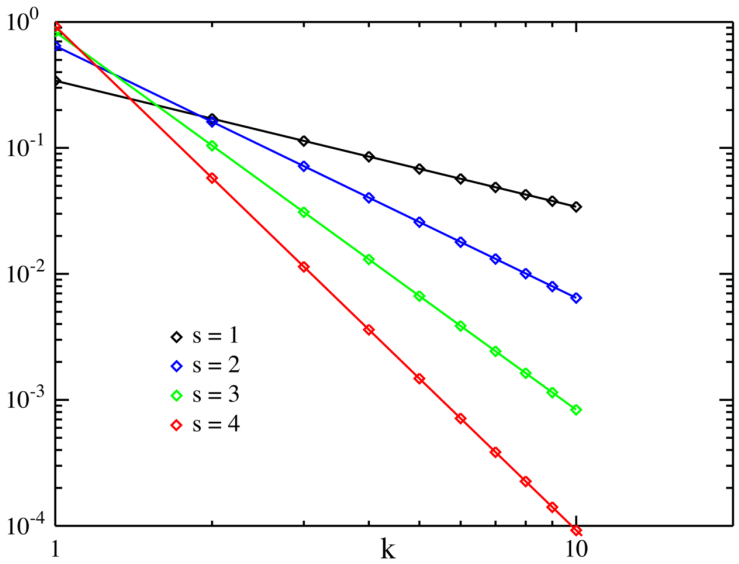
Modele grafów losowych

Model Erdősa-Rényi'ego
: Prawdopodobieństwo połączenia równe i niezależne od poprzednich połączeń.

Model Wattsa-Strogatza
: Krótkie średnie długości ścieżek i wysoki współczynnik klastrowania.

Model Barabásiego–Alberta
: Preferencyjne dołączanie.

Rozkład stopni wierzchołków



Rysunek : Prawo Zipfa

Prawo potęgowe:

$$P(k) = \frac{n_k}{n} \sim k^{-\gamma}, \qquad (1)$$

gdzie $P(k)$ to prawdopodobieństwo, że dany wierzchołek ma stopień równy k, n jest liczbą wierzchołków, a n_k liczbą wierzchołków o stopniu k.

Upraszczanie sieci

Usuwanie krawędzi o mniejszym znaczeniu.
Stosunek współczynnika spójności:

$$rk(V, E, E_R) = \frac{C(V, E, E_R)}{C(V, E)}, \tag{2}$$

gdzie V to zbiór wierzchołków, E to zbiór krawędzi, E_R to zbiór wierzchołków „do usunięcia", a $C(V, E, E_R)$ to spójność grafu po usunięciu z niego krawędzi E_R.

$$0 < rk < 1. \tag{3}$$

Szacowanie opinii

Miara szacowanej opinii:

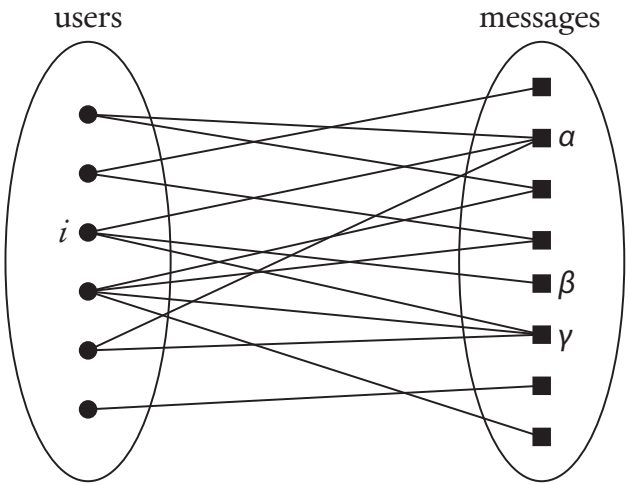
$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \ldots = \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \tag{4}$$

gdzie n jest odległością opinii od słowa kluczowego.
Przedział wartości miary:

$$-\frac{\pi^2}{3} \leq e_o \leq \frac{\pi^2}{3}, \tag{5}$$

wynika z możliwości położenia opinii zarówno przed jak i po słowie kluczowym.

Collaborative similarity



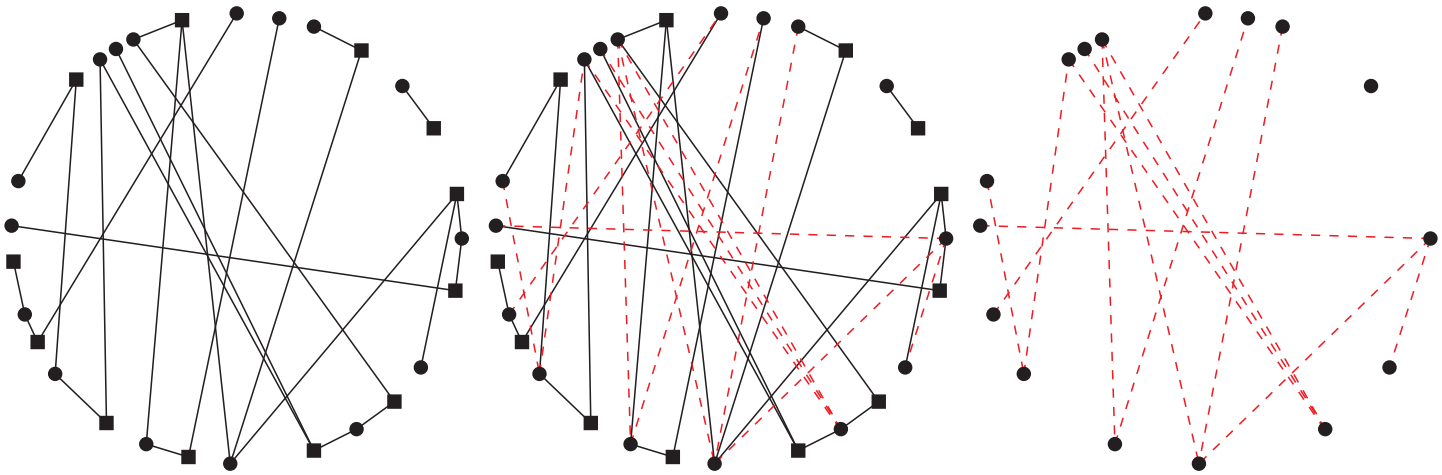
Rysunek : Sieć dwudzielna

$$d_{nn}(i) = \frac{d_\alpha + d_\beta + d_\gamma}{3} = \frac{7}{3}. \quad (6)$$

$$s_{\beta\gamma} = \frac{\Gamma_\beta \cap \Gamma_\gamma}{\Gamma_\beta \cup \Gamma_\gamma} = \frac{1}{3}. \quad (7)$$

$$C_u(i) = \frac{1}{k_i(k_i - 1)} \sum_{\alpha \neq \beta} s_{\alpha\beta}. \quad (8)$$

Motywy



(a) Krok 1. (b) Krok 2. (c) Krok 3.

Rysunek : Wyszukiwanie motywów.

Koła to użytkownicy, a kwadraty to słowa kluczowe. Linie ciągłe reprezentują słowa użyte przez użytkowników, a przerywane uzyskane połączenia między użytkownikami.

Podsumowanie

Efekty:

- Nie udało się znaleźć „nieszczerych” użytkowników.
- Zaproponowano miarę szacowania opinii oraz zaprezentowano sposób jej użycia.

Podsumowanie

Efekty:

- Nie udało się znaleźć „nieszczerych” użytkowników.
- Zaproponowano miarę szacowania opinii oraz zaprezentowano sposób jej użycia.

Co dalej?

- System rekomendujący na podstawie zmian w opiniach.
- Oszacowanie „idealnego” czasu na prowadzenie akcji marketingowej.

Koniec

Dziękuję za uwagę.