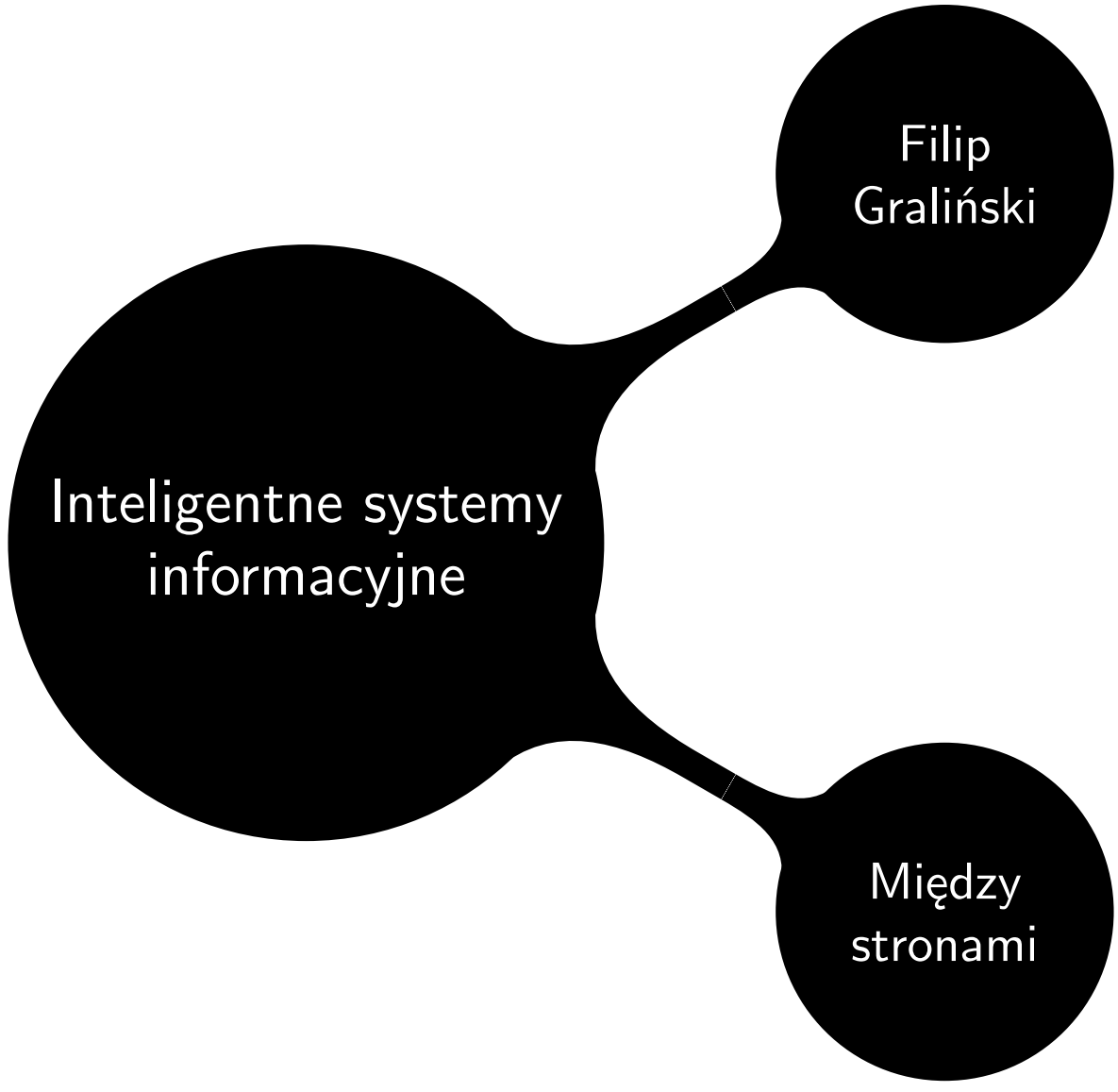
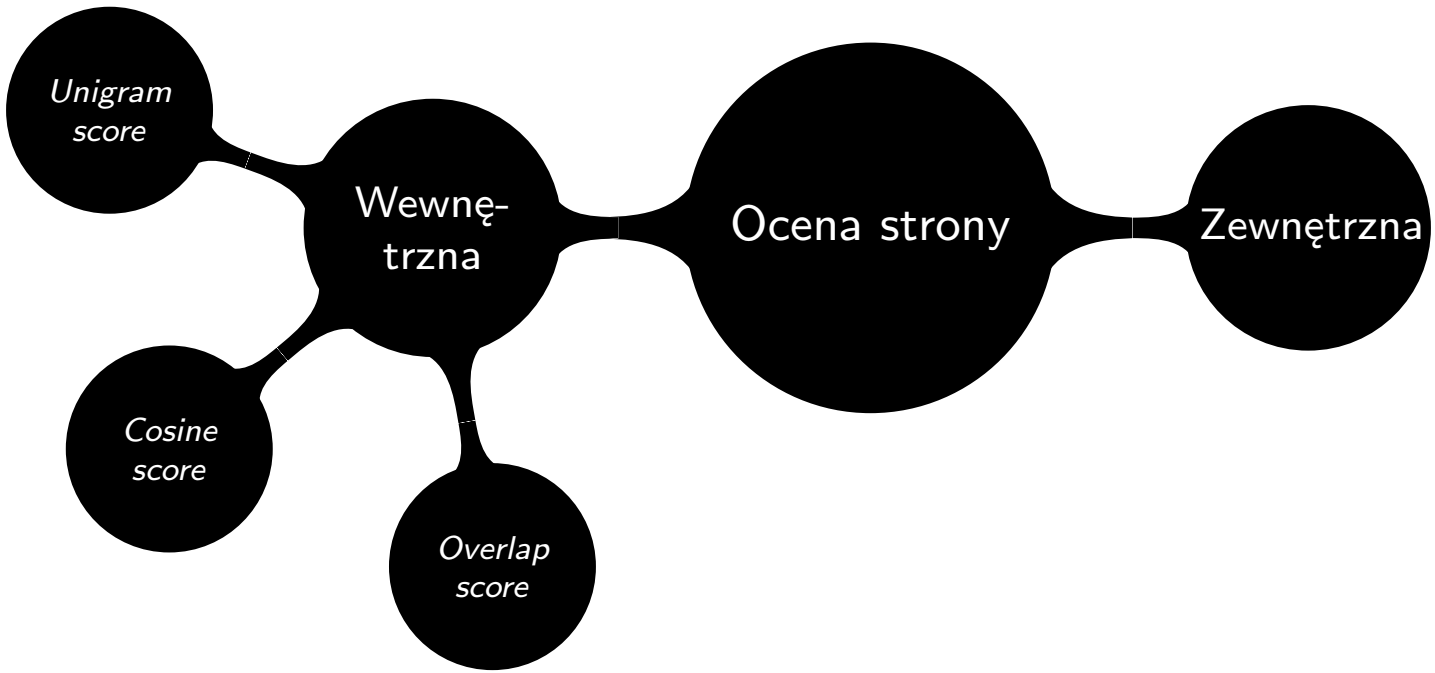
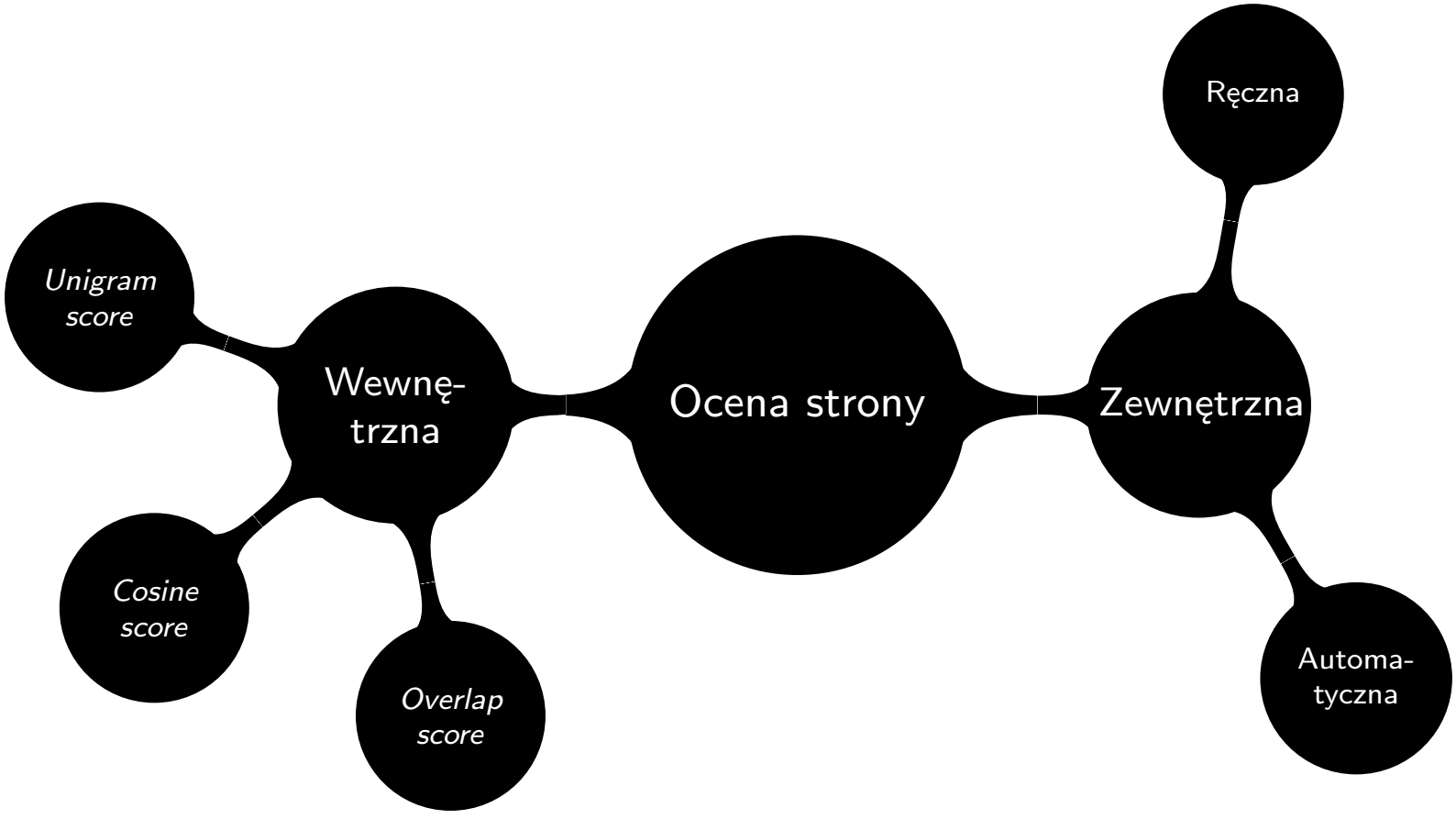
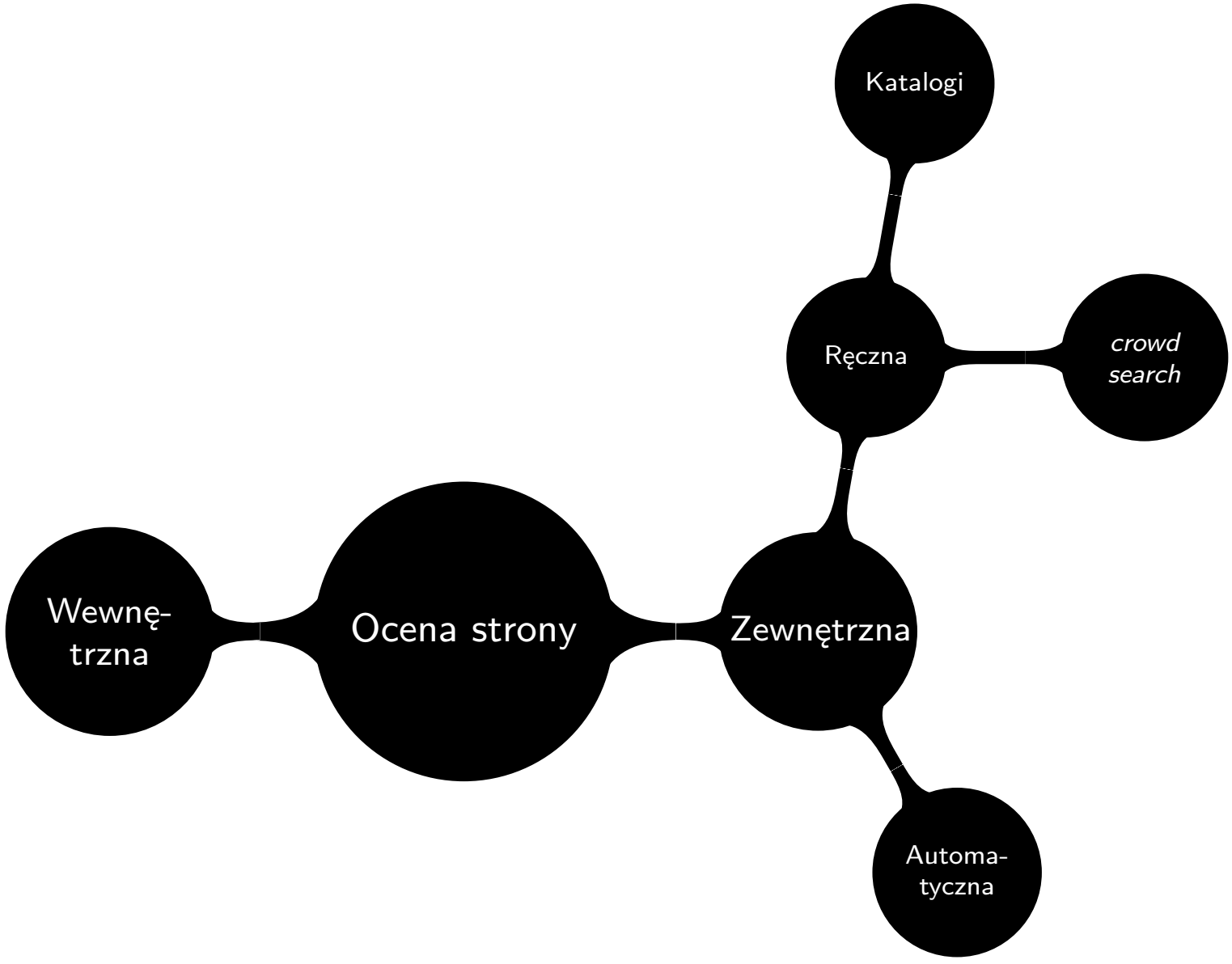
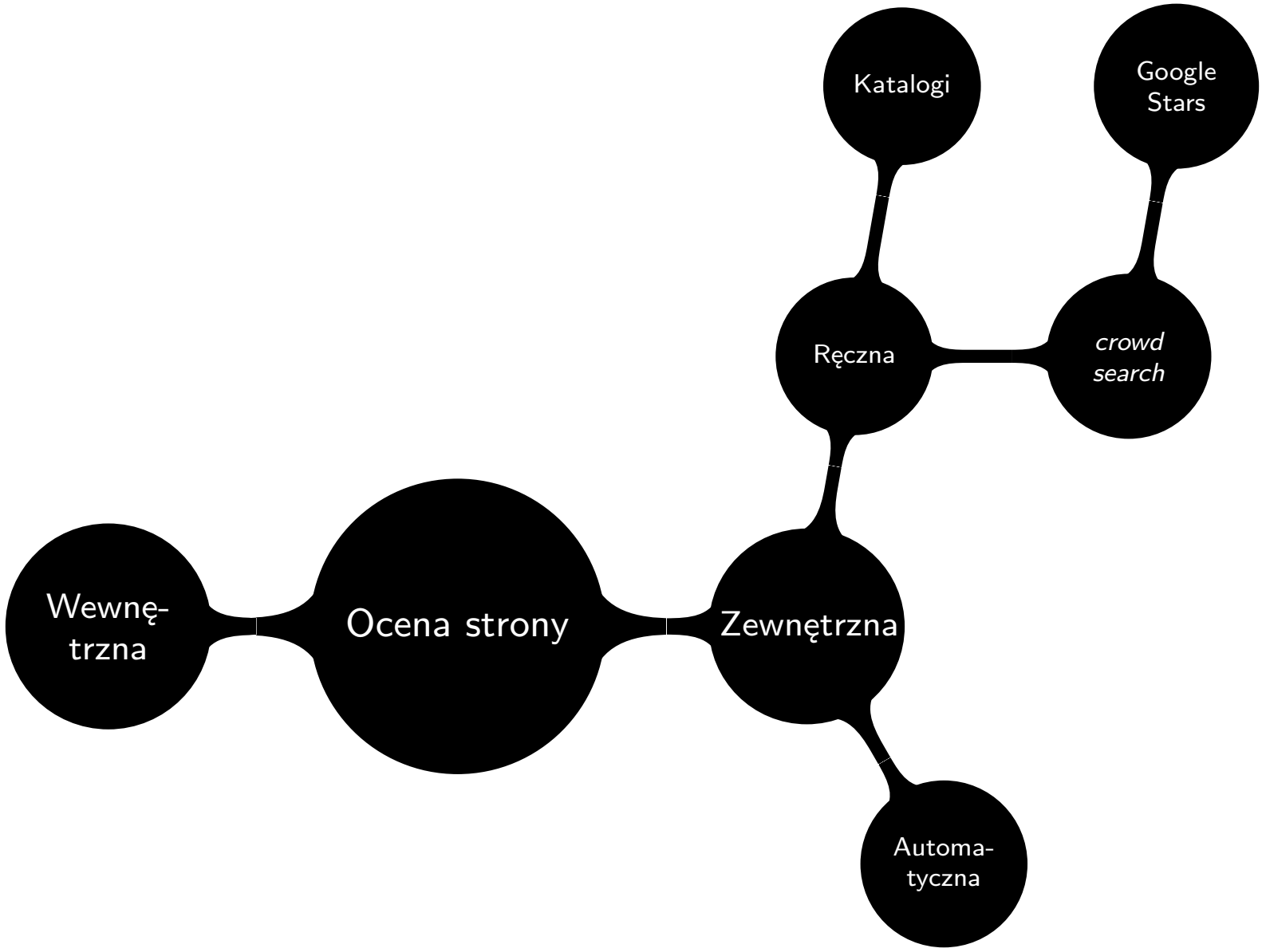

Filip
Graliński
Inteligentne systemy
informacyjne
Między
stronami

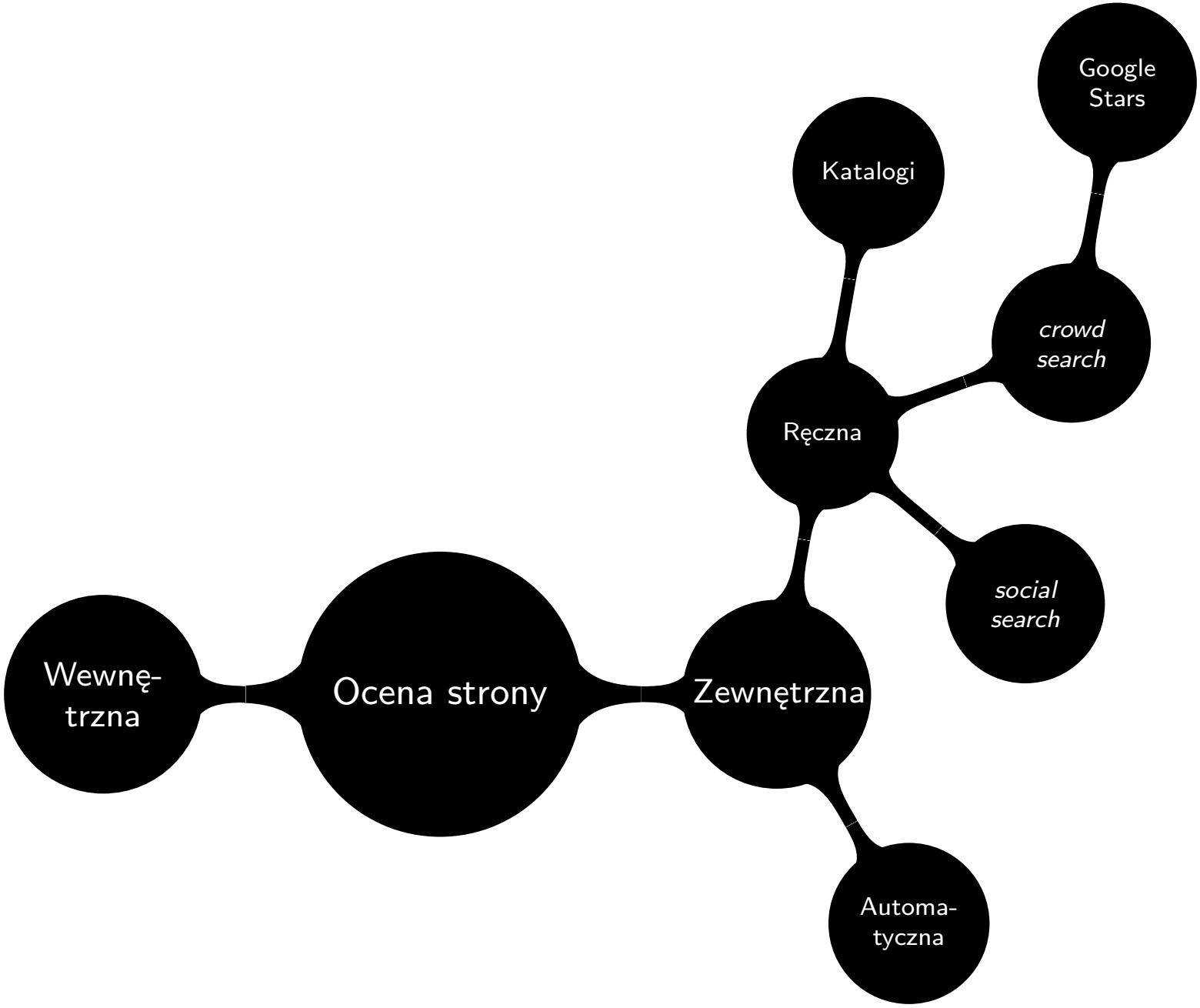
adversarial information retrieval =
„wyszukiwanie informacji
we wrogim środowisku” (spamerzy!)

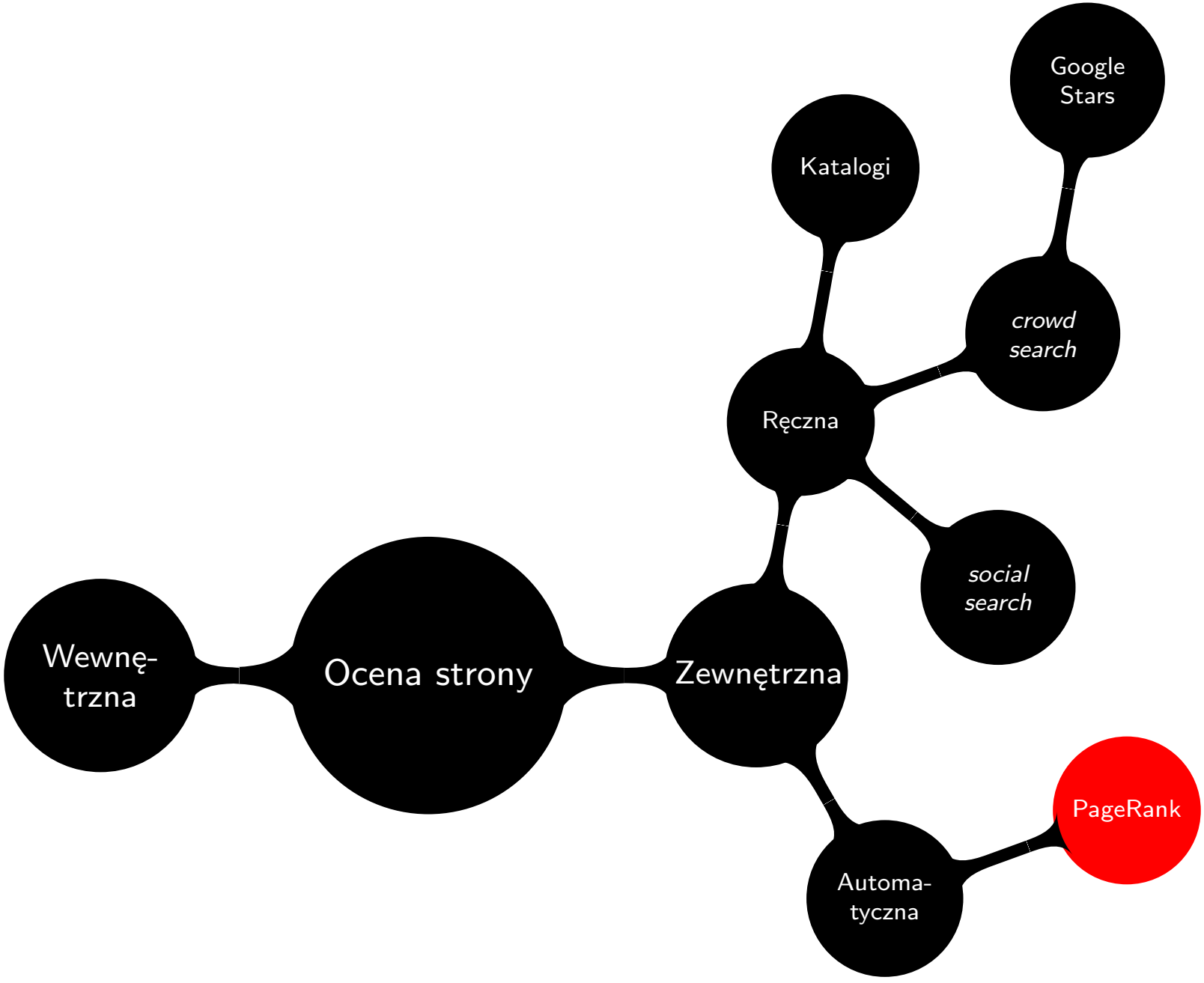

Unigram score
Wewnę-trzna
Cosine score
Overlap score
Ocena strony
Zewnętrzna


Ręczna
Unigram score
Wewnę-trzna
Ocena strony
Zewnętrzna
Cosine score
Overlap score
Automa-tyczna


Katalogi
Ręczna
crowd search
Wewnę-trzna
Ocena strony
Zewnętrzna
Automa-tyczna


Katalogi
Google Stars
Ręczna
crowd search
Wewnę-trzna
Ocena strony
Zewnętrzna
Automa-tyczna


Google Stars
Katalogi
crowd search
Ręczna
social search
Wewnę-trzna
Ocena strony
Zewnętrzna
Automa-tyczna


Google Stars
Katalogi
crowd search
Ręczna
social search
Wewnę-trzna
Ocena strony
Zewnętrzna
PageRank
Automa-tyczna

zasada PageRank

Dobry dokument to taki, do którego odsyła wiele dokumentów.

zasada PageRank

Dobry dokument to taki, do którego odsyła wiele innych dobrych dokumentów.

zasada PageRank

Dobry dokument to taki, do którego odsyła wiele innych dobrych dokumentów.

Model losowego internauty

zasada PageRank

Dobry dokument to taki, do którego odsyła wiele innych dobrych dokumentów.

Model losowego internauty

- losowy internauta zaczyna na losowo wybranej stronie

zasada PageRank

Dobry dokument to taki, do którego odsyła wiele innych dobrych dokumentów.

Model losowego internauty

- losowy internauta zaczyna na losowo wybranej stronie
- w każdym kroku losowy internauta przechodzi do kolejnej losowo wybranej strony

zasada PageRank

Dobry dokument to taki, do którego odsyła wiele innych dobrych dokumentów.

Model losowego internauty

- losowy internauta zaczyna na losowo wybranej stronie
- w każdym kroku losowy internauta przechodzi do kolejnej losowo wybranej strony
- jeśli bieżąca strona to ślepy zaułek (nie ma wychodzących linków), internauta *teleportuje się* na losową stronę

zasada PageRank

Dobry dokument to taki, do którego odsyła wiele innych dobrych dokumentów.

Model losowego internauty

- losowy internauta zaczyna na losowo wybranej stronie
- w każdym kroku losowy internauta przechodzi do kolejnej losowo wybranej strony
- jeśli bieżąca strona to ślepy zaułek (nie ma wychodzących linków), internauta *teleportuje się* na losową stronę
- jeśli bieżąca strona ma wychodzące linki

zasada PageRank

Dobry dokument to taki, do którego odsyła wiele innych dobrych dokumentów.

Model losowego internauty

- losowy internauta zaczyna na losowo wybranej stronie
- w każdym kroku losowy internauta przechodzi do kolejnej losowo wybranej strony
- jeśli bieżąca strona to ślepy zaułek (nie ma wychodzących linków), internauta *teleportuje się* na losową stronę
- jeśli bieżąca strona ma wychodzące linki
  - z prawdopodobieństwem α (zwykle $\alpha \approx 0.1$) internauta teleportuje się

zasada PageRank

Dobry dokument to taki, do którego odsyła wiele innych dobrych dokumentów.

Model losowego internauty

- losowy internauta zaczyna na losowo wybranej stronie
- w każdym kroku losowy internauta przechodzi do kolejnej losowo wybranej strony
- jeśli bieżąca strona to ślepy zaułek (nie ma wychodzących linków), internauta *teleportuje się* na losową stronę
- jeśli bieżąca strona ma wychodzące linki
  - z prawdopodobieństwem α (zwykle $\alpha \approx 0.1$) internauta teleportuje się
  - z prawdopodobieństwem $1 - \alpha$ przechodzi do strony, do której prowadzi losowo wybrany wychodzący link

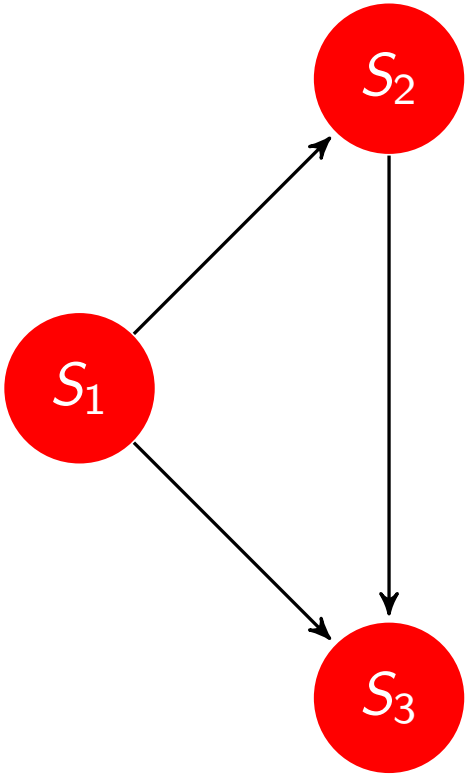
zasada PageRank

Dobry dokument to taki, do którego odsyła wiele innych dobrych dokumentów.

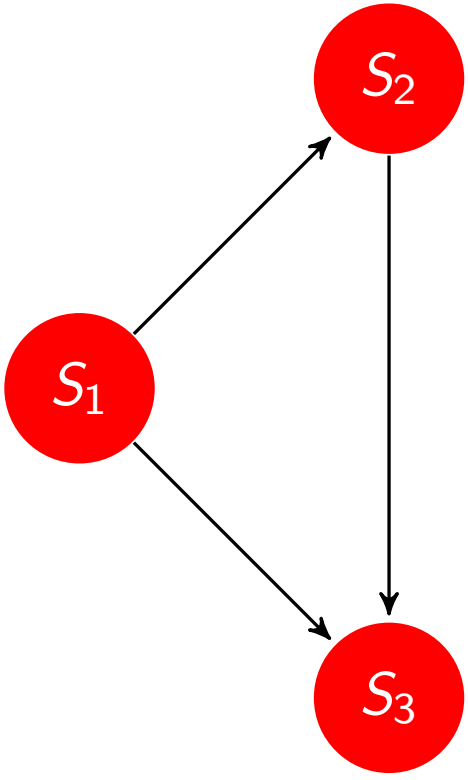
Model losowego internauty

- losowy internauta zaczyna na losowo wybranej stronie
- w każdym kroku losowy internauta przechodzi do kolejnej losowo wybranej strony
- jeśli bieżąca strona to ślepy zaułek (nie ma wychodzących linków), internauta *teleportuje się* na losową stronę
- jeśli bieżąca strona ma wychodzące linki
  - z prawdopodobieństwem α (zwykle $\alpha \approx 0.1$) internauta teleportuje się
  - z prawdopodobieństwem $1 - \alpha$ przechodzi do strony, do której prowadzi losowo wybrany wychodzący link
- PageRank strony to procent czasu, w jakim losowy internauta będzie na niej przebywał

macierz przejść



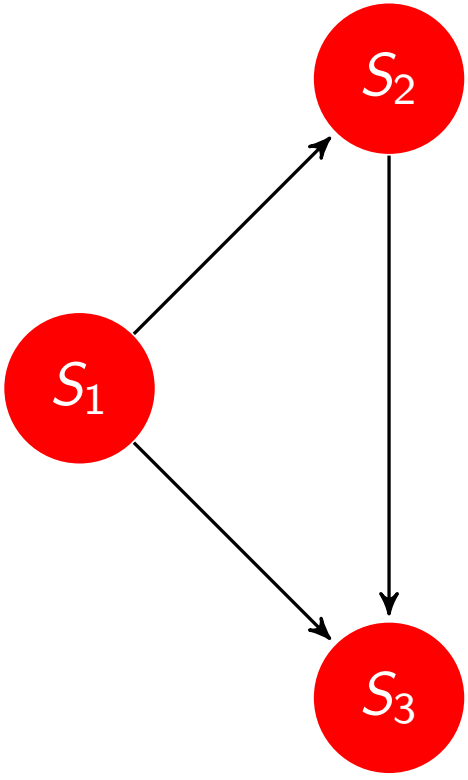
macierz przejść



$$\alpha = 0.1$$

$$M = \begin{bmatrix} 0.033 & 0.483 & 0.483 \\ 0.033 & 0.033 & 0.933 \\ 0.333 & 0.333 & 0.333 \end{bmatrix}$$

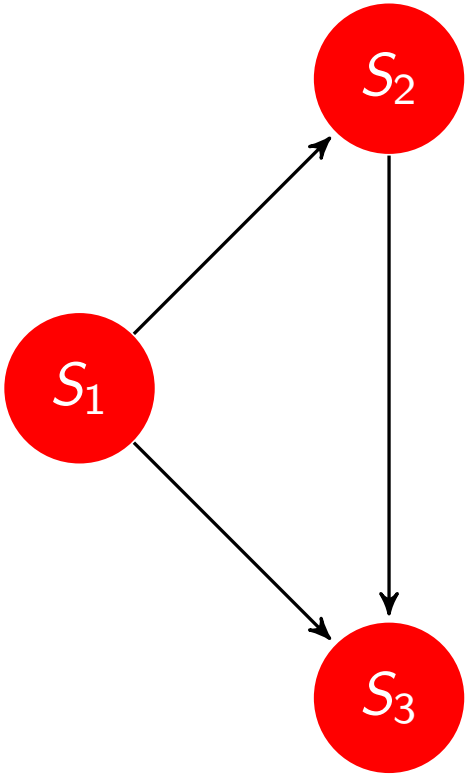
macierz przejść



$$\alpha = 0.1$$

$$M^2 = \begin{bmatrix} 0.177 & 0.192 & 0.627 \\ 0.312 & 0.327 & 0.357 \\ 0.132 & 0.282 & 0.582 \end{bmatrix}$$

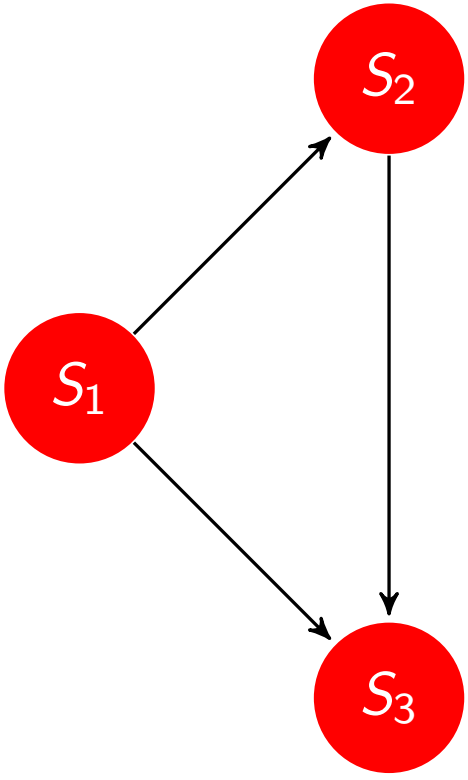
macierz przejść



$$\alpha = 0.1$$

$$M^3 = \begin{bmatrix} 0.221 & 0.301 & 0.474 \\ 0.140 & 0.280 & 0.575 \\ 0.207 & 0.267 & 0.521 \end{bmatrix}$$

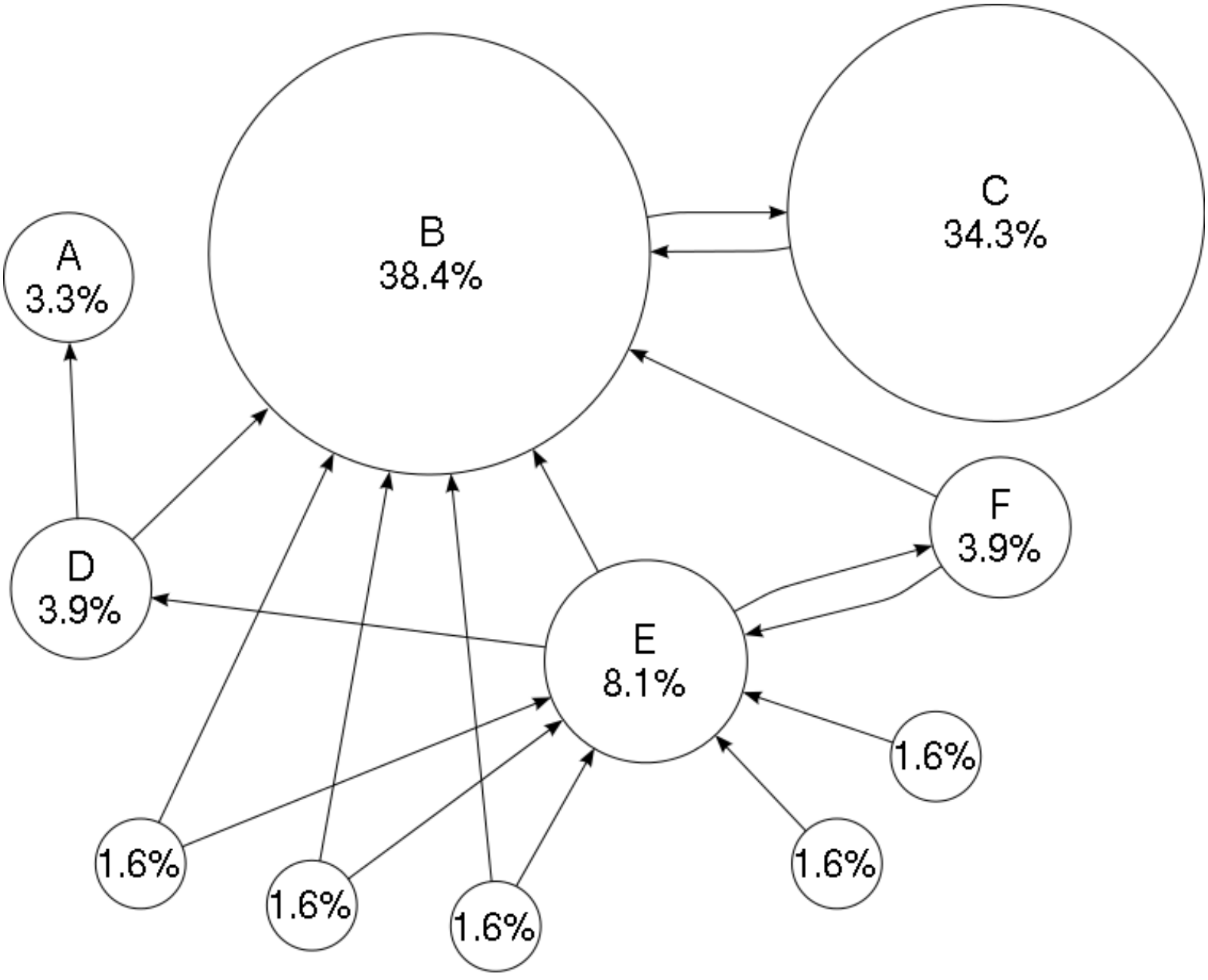
macierz przejść



$$\alpha = 0.1$$

$$M^{\infty} = \begin{bmatrix} 0.188 & 0.273 & 0.519 \\ 0.188 & 0.273 & 0.519 \\ 0.188 & 0.273 & 0.519 \end{bmatrix}$$

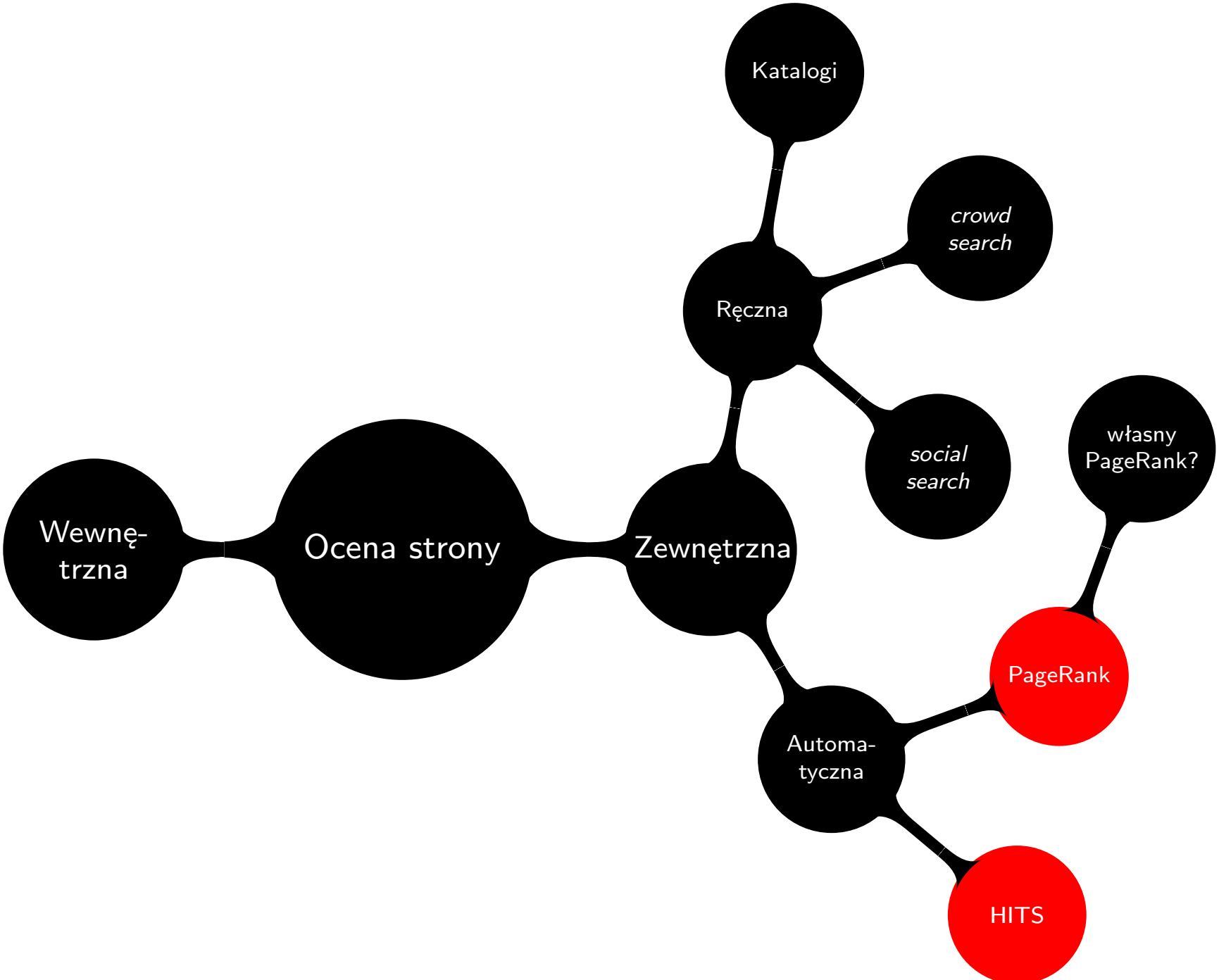
dlaczego C ma dużo większy PageRank niż E?

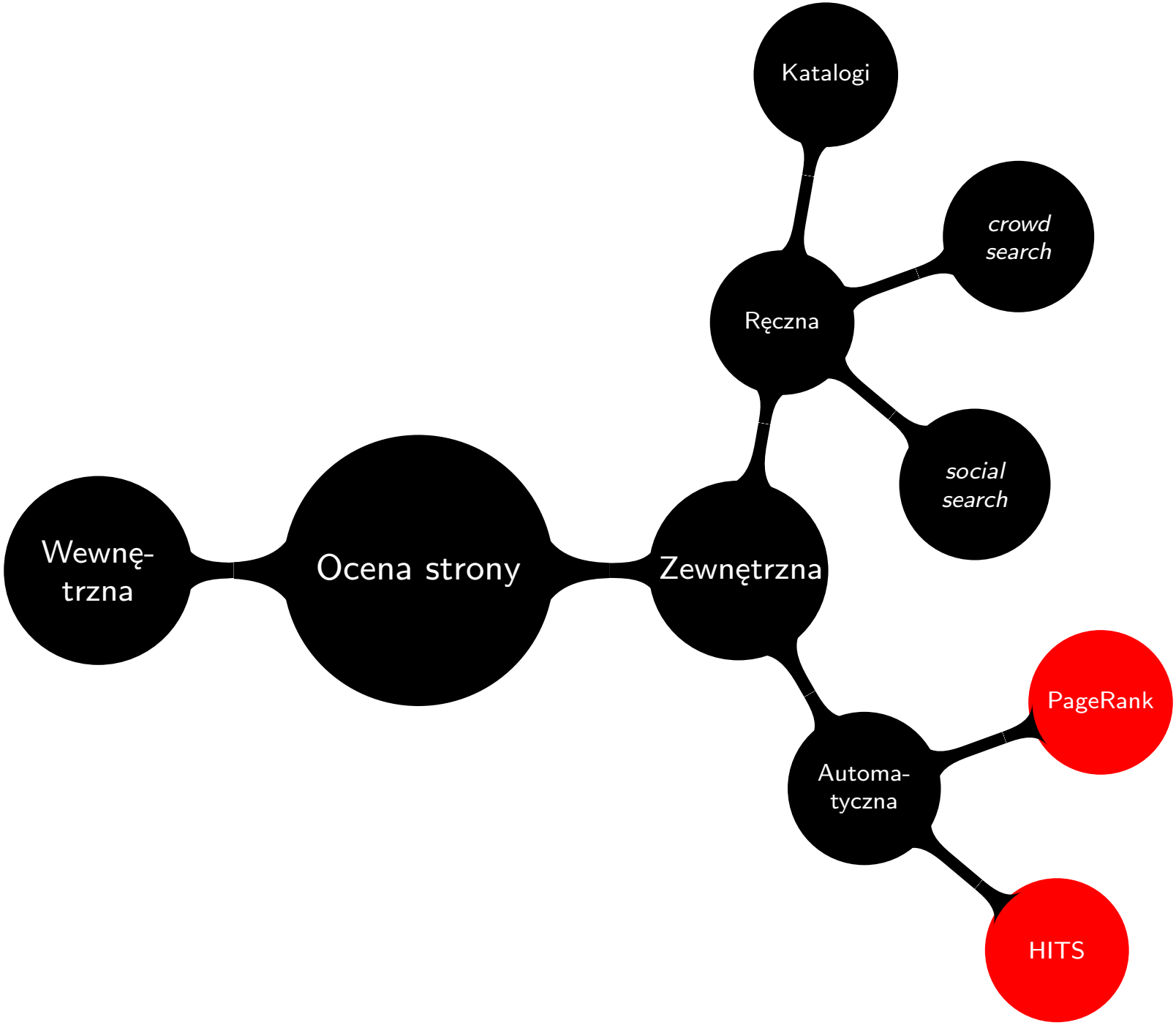


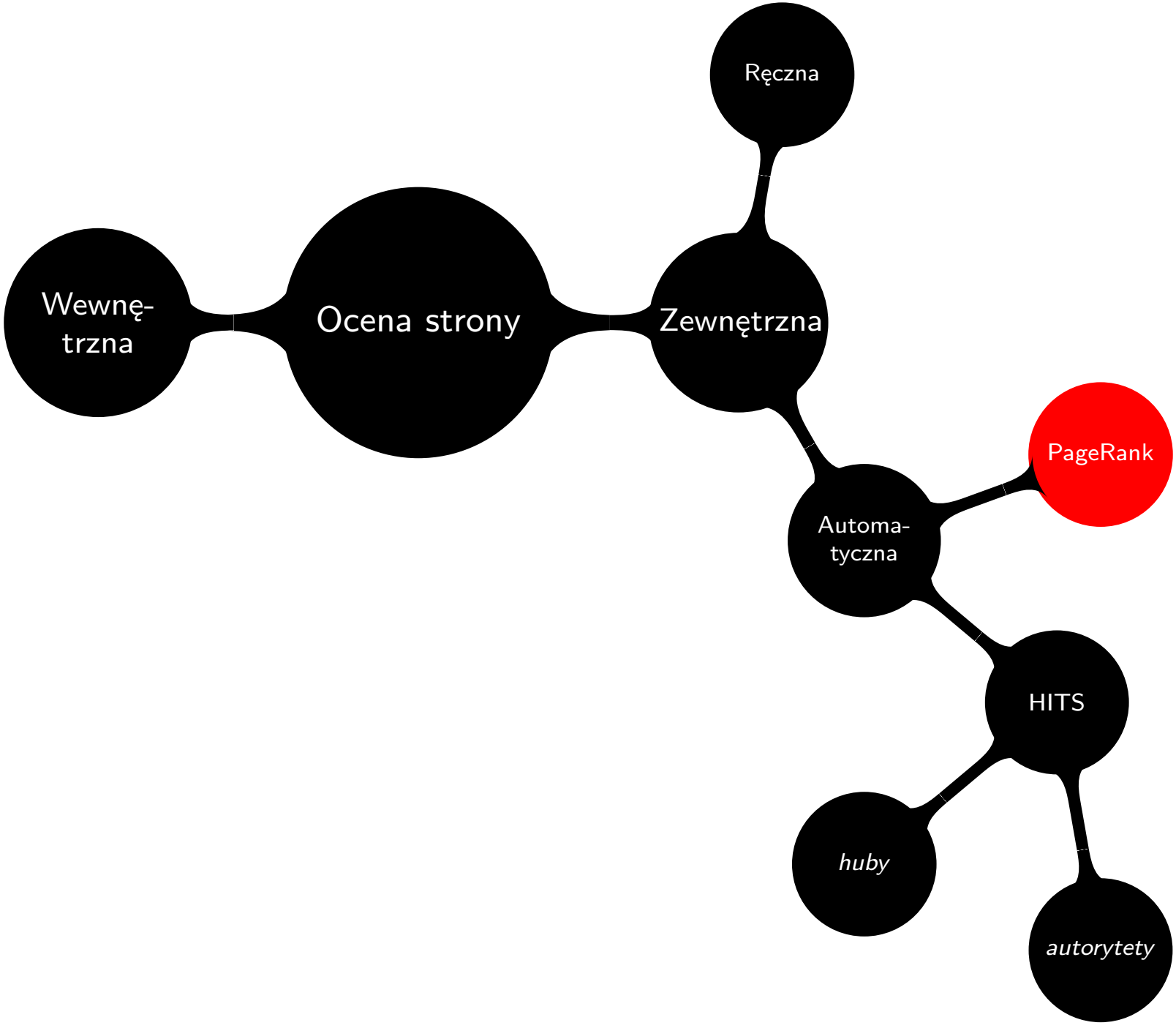
Projekt 3

Kto ma największy PersonRank?
Zastosować algorytm PageRank dla osób opisanych w polskiej Wikipedii:

- osoba A odsyła do osoby B, jeśli B pojawia się w biogramie A


Katalogi
crowd search
Ręczna
social search
własny PageRank?
Wewnę-trzna
Ocena strony
Zewnętrzna
PageRank
Automa-tyczna
HITS


Katalogi
crowd search
Ręczna
social search
Wewnę-trzna
Ocena strony
Zewnętrzna
PageRank
Automa-tyczna
HITS


Ręczna
Wewnę-
trzna
Ocena strony
Zewnętrzna
PageRank
Automa-
tyczna
HITS
huby
autorytety

HITS

Do dobrego autorytetu odsyła wiele dobrych hubów.

$$a(v) = \sum_{y \to v} h(y)$$

HITS

Do dobrego autorytetu odsyła wiele dobrych hubów.

$$a(v) = \sum_{y \to v} h(y)$$

Dobry hub odsyła do dobrych autorytetów.

$$h(v) = \sum_{v \to y} a(y)$$

HITS

Do dobrego autorytetu odsyła wiele dobrych hubów.

$$a(v) = \sum_{y \to v} h(y)$$

Dobry hub odsyła do dobrych autorytetów.

$$h(v) = \sum_{v \to y} a(y)$$

(Najork, Zaragoza, Taylor 2007):
PageRank $<$ HITS $<$

HITS

Do dobrego autorytetu odsyła wiele dobrych hubów.

$$a(v) = \sum_{y \to v} h(y)$$

Dobry hub odsyła do dobrych autorytetów.

$$h(v) = \sum_{v \to y} a(y)$$

(Najork, Zaragoza, Taylor 2007):
PageRank $<$ HITS $<$ zliczanie linków $+$ BM25F

Pozycjonowanie

Pozycjonowanie (ang. *SEO = Search Engine Optimization*)

Szare/czarne SEO:

- szpikowanie ukrytymi słowami kluczowymi

Pozycjonowanie

Pozycjonowanie (ang. *SEO = Search Engine Optimization*)

Szare/czarne SEO:

- szpikowanie ukrytymi słowami kluczowymi
- *cloaking* – inna wersja strony dla wyszukiwarki, inna dla zwykłego użytkownika

Pozycjonowanie

Pozycjonowanie (ang. *SEO = Search Engine Optimization*)

Szare/czarne SEO:

- szpikowanie ukrytymi słowami kluczowymi
- *cloaking* – inna wersja strony dla wyszukiwarki, inna dla zwykłego użytkownika
- spamowanie komentarzy/stron wiki
- spamowanie list typu „odsyłają do nas”

Pozycjonowanie

Pozycjonowanie (ang. *SEO* = *Search Engine Optimization*)

Szare/czarne SEO:

- szpikowanie ukrytymi słowami kluczowymi
- *cloaking* – inna wersja strony dla wyszukiwarki, inna dla zwykłego użytkownika
- spamowanie komentarzy/stron wiki
- spamowanie list typu „odsyłają do nas”
- tworzenie „klik” powiązanych stron, źródła tekstu:

Pozycjonowanie

Pozycjonowanie (ang. *SEO = Search Engine Optimization*)

Szare/czarne SEO:

- szpikowanie ukrytymi słowami kluczowymi
- *cloaking* – inna wersja strony dla wyszukiwarki, inna dla zwykłego użytkownika
- spamowanie komentarzy/stron wiki
- spamowanie list typu „odsyłają do nas”
- tworzenie „klik” powiązanych stron, źródła tekstu:
  - pobrany z innych stron WWW

Pozycjonowanie

Pozycjonowanie (ang. *SEO = Search Engine Optimization*)

Szare/czarne SEO:

- szpikowanie ukrytymi słowami kluczowymi
- *cloaking* – inna wersja strony dla wyszukiwarki, inna dla zwykłego użytkownika
- spamowanie komentarzy/stron wiki
- spamowanie list typu „odsyłają do nas”
- tworzenie „klik” powiązanych stron, źródła tekstu:
  - pobrany z innych stron WWW
  - napisany przez tanich „copywriterów”

Pozycjonowanie

Pozycjonowanie (ang. *SEO = Search Engine Optimization*)

Szare/czarne SEO:

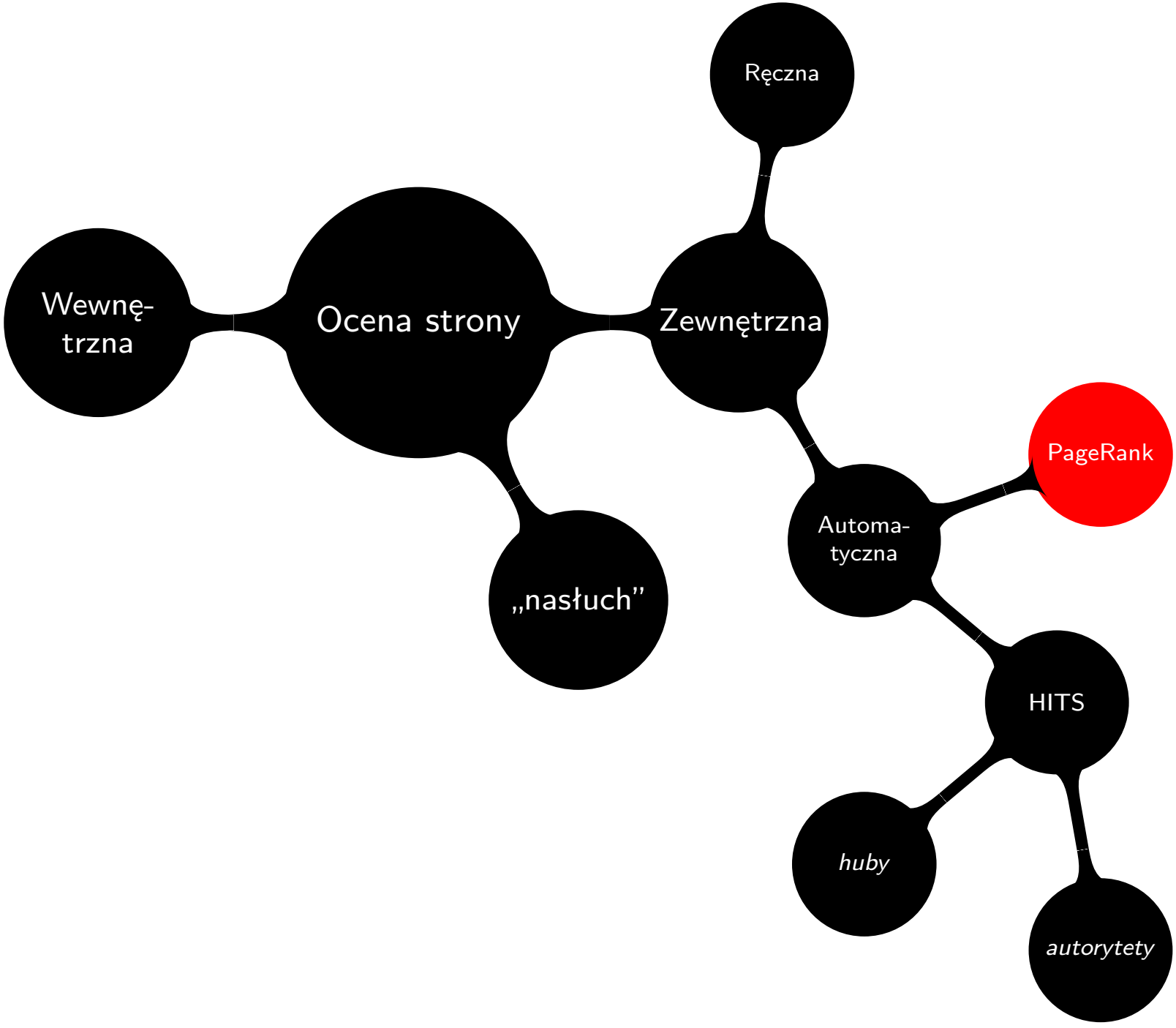
- szpikowanie ukrytymi słowami kluczowymi
- *cloaking* – inna wersja strony dla wyszukiwarki, inna dla zwykłego użytkownika
- spamowanie komentarzy/stron wiki
- spamowanie list typu „odsyłają do nas”
- tworzenie „klik” powiązanych stron, źródła tekstu:
  - pobrany z innych stron WWW
  - napisany przez tanich „copywriterów”
  - wygenerowany automatycznie

Pozycjonowanie

Pozycjonowanie (ang. *SEO = Search Engine Optimization*)

Szare/czarne SEO:

- szpikowanie ukrytymi słowami kluczowymi
- *cloaking* – inna wersja strony dla wyszukiwarki, inna dla zwykłego użytkownika
- spamowanie komentarzy/stron wiki
- spamowanie list typu „odsyłają do nas”
- tworzenie „klik” powiązanych stron, źródła tekstu:
  - pobrany z innych stron WWW
  - napisany przez tanich „copywriterów”
  - wygenerowany automatycznie
- kupowanie linków na stronach o dużym PageRanku


Ręczna
Wewnę-
trzna
Ocena strony
Zewnętrzna
PageRank
Automa-
tyczna
„nasłuch”
HITS
huby
autorytety

model intencjonalnego internauty

Jak wyszukiwarka może dowiedzieć, jakie strony internauci *naprawdę* odwiedzają?

- śledzenie klikanych linków:

```
<a href="http://www.pogotowiekrawieckie.pl/index.php?id=2" class=l
   onmousedown="return clk(this.href,'','','res','2','')">
<em>POGOTOWIE KRAWIECKIE</em></a>
```

model intencjonalnego internauty

Jak wyszukiwarka może dowiedzieć, jakie strony internauci *naprawdę* odwiedzają?

- śledzenie klikanych linków:

```
<a href="http://www.pogotowiekrawieckie.pl/index.php?id=2" class=l
   onmousedown="return clk(this.href,'','','res','2','')">
<em>POGOTOWIE KRAWIECKIE</em></a>
```

- Google Toolbar

model intencjonalnego internauty

Jak wyszukiwarka może dowiedzieć, jakie strony internauci *naprawdę* odwiedzają?

- śledzenie klikanych linków:

```
<a href="http://www.pogotowiekrawieckie.pl/index.php?id=2" class=l
   onmousedown="return clk(this.href,'','','res','2','')">
<em>POGOTOWIE KRAWIECKIE</em></a>
```

- Google Toolbar
- Google Analytics

model intencjonalnego internauty

Jak wyszukiwarka może dowiedzieć, jakie strony internauci *naprawdę* odwiedzają?

- śledzenie klikanych linków:

```
<a href="http://www.pogotowiekrawieckie.pl/index.php?id=2" class=l
   onmousedown="return clk(this.href,'','','res','2','')">
<em>POGOTOWIE KRAWIECKIE</em></a>
```

- Google Toolbar
- Google Analytics
- Google Chrome