AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Lingwistyczny system definicyjny wykorzystujący korpusy tekstów oraz zasoby internetowe.

Autor: Mariusz Sasko

Promotor: dr Adrian Horzyk

Plan prezentacji

1. Wstęp
2. Cele pracy
3. Rozwiązanie
    - 3.1. Robot internetowy
    - 3.2. Diagramy LHG
    - 3.3. Algorytm ekstrakcji informacji
4. Prezentacja wyniku
    - 4.1. Prezentacja aplikacji
    - 4.2. Prezentacja definicji hasła „komputer kwantowy”
5. Wnioski
6. Podsumowanie

Wstęp

- Powstanie i dynamiczny rozwój Internetu w XX wieku oraz wszechobecna komputeryzacja przyspieszyły rozwój społeczeństwa informacyjnego, w którym głównym towarem mającym największą wartość jest informacja.

- Człowiek uzyskał dostęp do narzędzi, za pomocą których może szybko pozyskiwać nowe informacje, selekcjonować je, analizować, przetwarzać, zarządzać nimi oraz przekazywać je innym ludziom na bardzo duże odległości w bardzo krótkim czasie.

- Ogromna ilość informacji zgromadzonych w Internecie spowodowała konieczność stworzenia wyszukiwarek, które umożliwiają dostęp do pewnych stron, bez znajomości ich adresów, na podstawie wprowadzonej kwerendy.

- Ciągle jednak prowadzone są badania nad szybszymi sposobami dostępu do informacji, uzyskanymi z sieci Internet.

Cele pracy

Celem pracy było skonstruowanie systemu posiadającego wyspecjalizowany algorytm ekstrakcji informacji z tekstu, w celu stworzenia definicji dowolnego znaczenia.

Założenia:

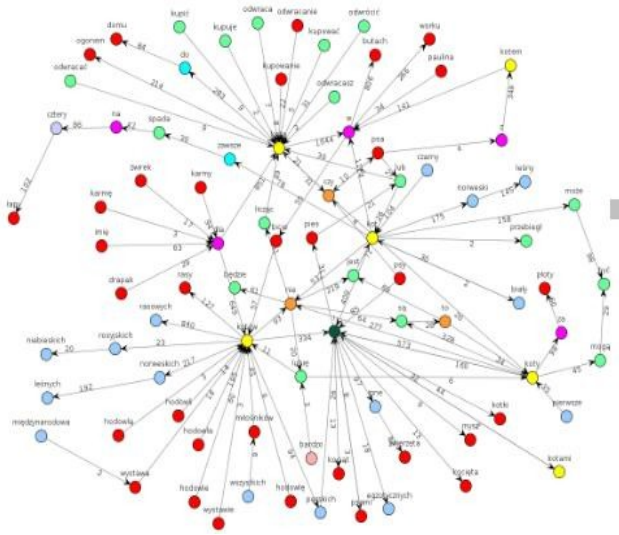
- Budowa uniwersalnego systemu, działającego dla każdego znaczenia, o którym można znaleźć informacje w Internecie.
- System powinien samodzielnie pobierać z Internetu teksty, zawierające istotne dane na temat definiowanego hasła, w postaci ciągów wyrazów i na ich podstawie budować definicję.
- Praca wykonywana dla języka polskiego.

Rozwiązanie - schemat działania algorytmu

Google

Zasoby internetowe

Robot internetowy



Graf LHG

Hasło: systemy wizyjne

Jakie to jest? - zestaw przymiotników opisujących d

1. **współpracujący** - współpracującego z systemem wizyjny
2. **komputerowy** - komputerowych systemów wizyjnych(22)
3. **skomplikowany** - skomplikowanych systemach wizyjnych
4. **elastyczny** - elastyczne systemy wizyjne(4),
5. **niemożliwy** - analiza danych w systemach wizyjnych była
6. **zaawansowany** - zaawansowane systemy wizyjne(24), za
7. **kompletny** - kompletne systemy wizyjne(10),
8. **wyposażony** - wyposażone w system wizyjny(11),
9. **przemysłowy** - przemysłowe systemy wizyjne(35),
10. **badany** - badanych przez systemy wizyjne(2),

Definicja

Robot internetowy

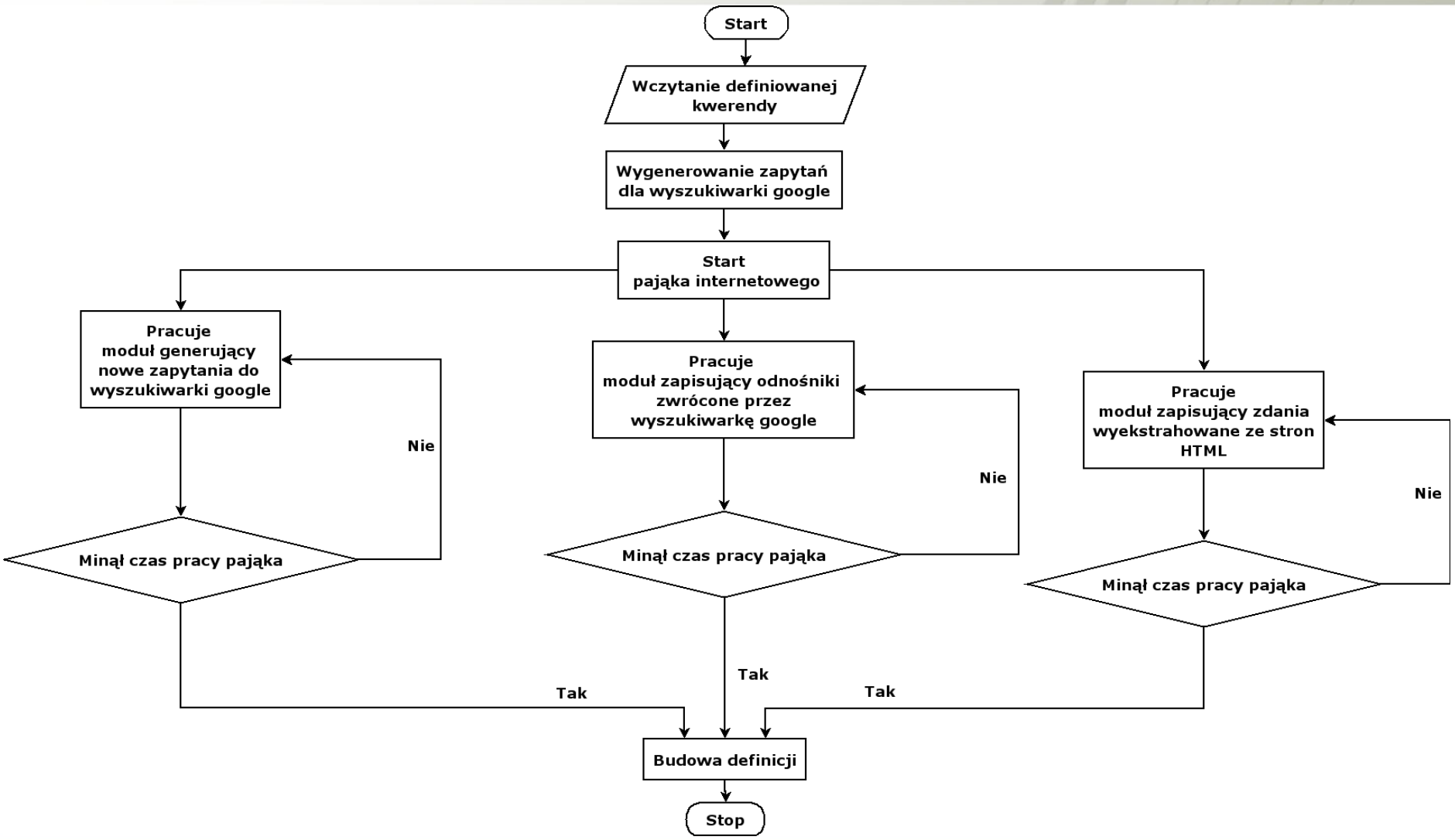
Roboty internetowe, zwane także pająkami internetowymi, stanowią podstawowy element działania wszystkich współczesnych wyszukiwarek.

- Przechodzą one po stronach internetowych, zbierając informacje. Kiedy napotkają nową stronę, której jeszcze nie zindeksowały, najczęściej pobierają jej treść bez obrazków oraz zapisują znalezione odnośniki i przechodzą do analizy kolejnych stron.

Roboty internetowe spełniają różne funkcje. Najczęściej wykorzystuje się je do:

- wyszukiwania nowych stron internetowych dla wyszukiwarek,
- analizy kodu strony,
- gromadzenia informacji o stronach,
- analizy zmian zachodzących na stronach.

Robot internetowy – algorytm działania



Pozyskiwanie informacji z Internetu

Głównym zadaniem pająka internetowego było wyszukiwanie zdań zawierających definiowaną frazę w dowolnej formie fleksyjnej, ze stron pod odnośnikami zwróconymi przez wyszukiwarkę google.com oraz zapisywanie ich na diagramie LHG.

Problemy:

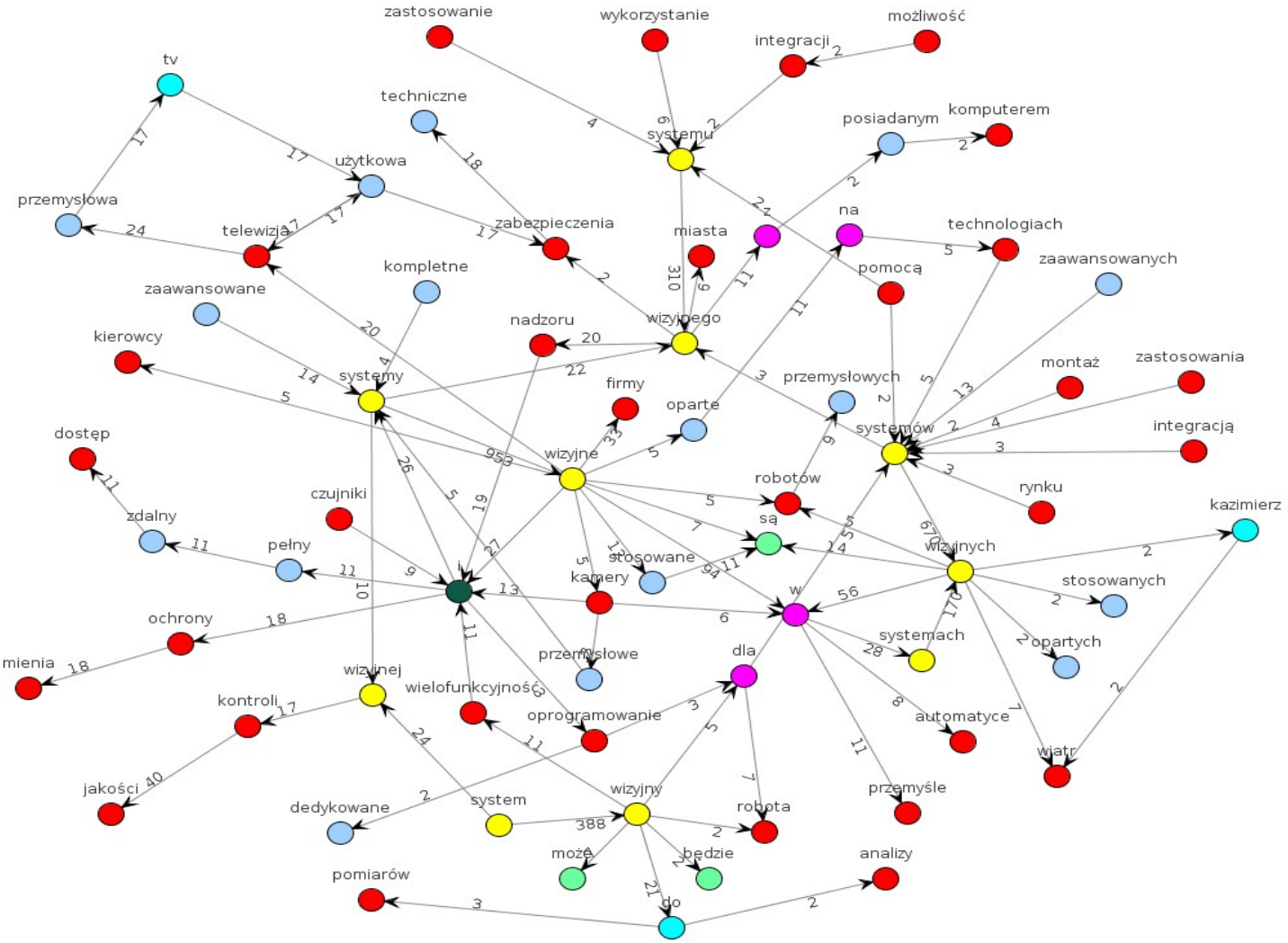
- Wyszukiwanie tekstów zawierających definiowane hasło w dowolnej formie fleksyjnej, wygenerowanej za pomocą biblioteki lingwistycznej CLP, z której skorzystano dzięki uprzejmości p. dr inż Marka Gajęckiego z krakowskiej Grupy Lingwistyki Komputerowej i opracowanej przez zakład prof dr hab. Wiesława Lubaszewskiego.
- Ekstrakcja zdań. Wykorzystany został specjalnie skonstruowany mechanizm, zabezpieczający przed pomieszaniem słów, po usunięciu znaczników HTML.
- Zapisywanie wyszukanych ciągów słów na diagramie LHG.

Grafy Przyzwyczajeń Lingwistycznych (LHG)

Grafy LHG są uproszczoną wersją Grafów Lingwistycznych Neuroasocjacji Semantycznych (GLAS), które zostały zaprojektowane przez promotora pracy dra Adriana Horzyka.

- Najważniejszą cechą grafów LHG jest utrzymywanie kontekstu całej wypowiedzi.
- Bardzo ważną część grafu stanowią krawędzie, które powstają w wyniku parsowania zdań. Stanowią one odzwierciedlenie połączeń, jakie mogą występować pomiędzy słowami i są one nazywane asocjacjami.
- Węzłami na grafie LHG są słowa. Każda forma fleksyjna może wystąpić tylko jeden raz.
- Grafy LHG gromadzą informacje ilościowe, dotyczące częstość występowania słów i asocjacji.

Fragment grafu LHG dla hasła „systemy wizyjne”



Ekstrakcja informacji

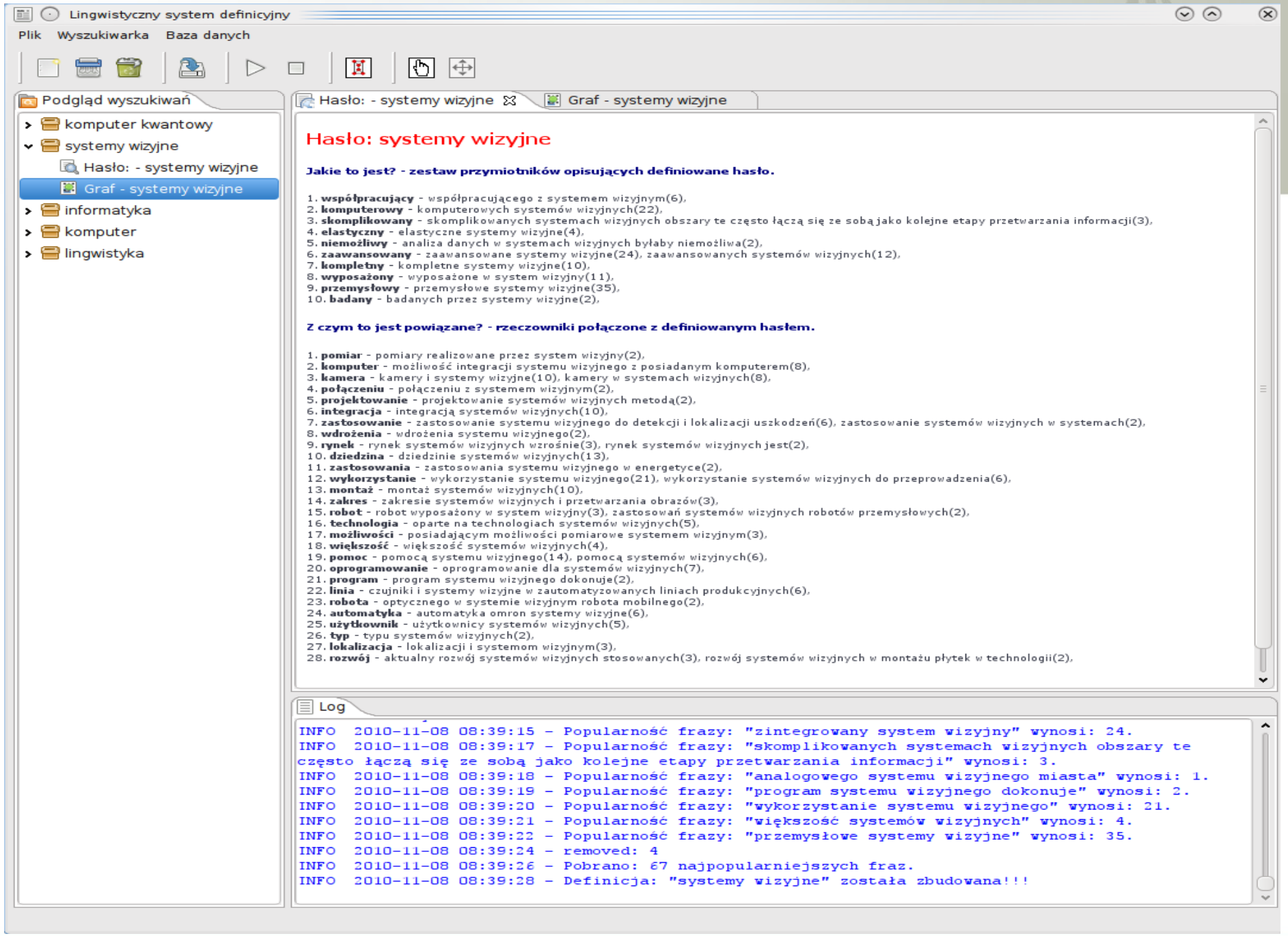
Proces ten ma na celu wyszukanie ciągów słów w postaci fraz, na podstawie danych ilościowych zgromadzonych na diagramie LHG.

- Zaproponowany algorytm analizuje kolejno częstość występowania asocjacji w sąsiedztwie definiowanego hasła i tworzy z nich ciągi słów.

- Wykorzystana została wyszukiwarka google.com do sprawdzenia popularności występowania całej frazy w Internecie i na bazie uzyskanego wyniku, odrzucane są te najmniej popularne.

- Uzyskane frazy są analizowane i wyszukiwane są przymiotniki, czasowniki i rzeczowniki, które służą do budowy definicji.

Prezentacja aplikacji


Lingwistyczny system definicyjny
Plik Wyszukiwarka Baza danych
Podgląd wyszukiwań
komputer kwantowy
systemy wizyjne
Hasło: - systemy wizyjne
Graf - systemy wizyjne
informatyka
komputer
lingwistyka
Hasło: - systemy wizyjne
Graf - systemy wizyjne
Hasło: systemy wizyjne
Jakie to jest? - zestaw przymiotników opisujących definiowane hasło.
1. współpracujący - współpracującego z systemem wizyjnym(6),
2. komputerowy - komputerowych systemów wizyjnych(22),
3. skomplikowany - skomplikowanych systemach wizyjnych obszary te często łączą się ze sobą jako kolejne etapy przetwarzania informacji(3),
4. elastyczny - elastyczne systemy wizyjne(4),
5. niemożliwy - analiza danych w systemach wizyjnych byłaby niemożliwa(2),
6. zaawansowany - zaawansowane systemy wizyjne(24), zaawansowanych systemów wizyjnych(12),
7. kompletny - kompletne systemy wizyjne(10),
8. wyposażony - wyposażone w system wizyjny(11),
9. przemysłowy - przemysłowe systemy wizyjne(35),
10. badany - badanych przez systemy wizyjne(2),
Z czym to jest powiązane? - rzeczowniki połączone z definiowanym hasłem.
1. pomiar - pomiary realizowane przez system wizyjny(2),
2. komputer - możliwość integracji systemu wizyjnego z posiadanym komputerem(8),
3. kamera - kamery i systemy wizyjne(10), kamery w systemach wizyjnych(8),
4. połączeniu - połączeniu z systemem wizyjnym(2),
5. projektowanie - projektowanie systemów wizyjnych metodą(2),
6. integracja - integracją systemów wizyjnych(10),
7. zastosowanie - zastosowanie systemu wizyjnego do detekcji i lokalizacji uszkodzeń(6), zastosowanie systemów wizyjnych w systemach(2),
8. wdrożenia - wdrożenia systemu wizyjnego(2),
9. rynek - rynek systemów wizyjnych wzrośnie(3), rynek systemów wizyjnych jest(2),
10. dziedzina - dziedzinie systemów wizyjnych(13),
11. zastosowania - zastosowania systemu wizyjnego w energetyce(2),
12. wykorzystanie - wykorzystanie systemu wizyjnego(21), wykorzystanie systemów wizyjnych do przeprowadzenia(6),
13. montaż - montaż systemów wizyjnych(10),
14. zakres - zakresie systemów wizyjnych i przetwarzania obrazów(3),
15. robot - robot wyposażony w system wizyjny(3), zastosowań systemów wizyjnych robotów przemysłowych(2),
16. technologia - oparte na technologiach systemów wizyjnych(5),
17. możliwości - posiadającym możliwości pomiarowe systemem wizyjnym(3),
18. większość - większość systemów wizyjnych(4),
19. pomoc - pomocą systemu wizyjnego(14), pomocą systemów wizyjnych(6),
20. oprogramowanie - oprogramowanie dla systemów wizyjnych(7),
21. program - program systemu wizyjnego dokonuje(2),
22. linia - czujniki i systemy wizyjne w zautomatyzowanych liniach produkcyjnych(6),
23. robota - optycznego w systemie wizyjnym robota mobilnego(2),
24. automatyka - automatyka omron systemy wizyjne(6),
25. użytkownik - użytkownicy systemów wizyjnych(5),
26. typ - typu systemów wizyjnych(2),
27. lokalizacja - lokalizacji i systemom wizyjnym(3),
28. rozwój - aktualny rozwój systemów wizyjnych stosowanych(3), rozwój systemów wizyjnych w montażu płytek w technologii(2),
Log
INFO 2010-11-08 08:39:15 - Popularność frazy: "zintegrowany system wizyjny" wynosi: 24.
INFO 2010-11-08 08:39:17 - Popularność frazy: "skomplikowanych systemach wizyjnych obszary te często łączą się ze sobą jako kolejne etapy przetwarzania informacji" wynosi: 3.
INFO 2010-11-08 08:39:18 - Popularność frazy: "analogowego systemu wizyjnego miasta" wynosi: 1.
INFO 2010-11-08 08:39:19 - Popularność frazy: "program systemu wizyjnego dokonuje" wynosi: 2.
INFO 2010-11-08 08:39:20 - Popularność frazy: "wykorzystanie systemu wizyjnego" wynosi: 21.
INFO 2010-11-08 08:39:21 - Popularność frazy: "większość systemów wizyjnych" wynosi: 4.
INFO 2010-11-08 08:39:22 - Popularność frazy: "przemysłowe systemy wizyjne" wynosi: 35.
INFO 2010-11-08 08:39:24 - removed: 4
INFO 2010-11-08 08:39:26 - Pobrano: 67 najpopularniejszych fraz.
INFO 2010-11-08 08:39:28 - Definicja: "systemy wizyjne" została zbudowana!!!

Prezentacja wyniku

Definicja w postaci listy słów z uzasadnieniem w postaci wyszukanych fraz, najpopularniejszych dla danego słowa.

Definicja została podzielona na cztery grupy:

- przymiotniki opisujące definiowane hasło,
- czasowniki będące orzeczeniami dla podmiotu będącego definiowanym hasłem,
- czasowniki nie będące orzeczeniami dla podmiotu będącego definiowanym hasłem,
- rzeczowniki związane z definiowanym hasłem.

Prezentacja wyniku

Hasło: komputer kwantowy

Jakie to jest? - zestaw przymiotników opisujących definiowane hasło.

1. **komercyjny** - określone jako pierwszy komercyjny komputer kwantowy(2),
2. **domowy** - domowe komputery kwantowe(4),
3. **zwany** - zwanych komputerami kwantowymi(2),
4. **możliwy** - komputerach kwantowych możliwe byłoby wykonanie tych operacji w bardziej realnym okresie(2),
5. **mały** - mały komputer kwantowy miałby(2),

Co można z tym robić? Jaki można mieć na to wpływ?

1. **zbudować** - zbudować komputer kwantowy szybki(2),

Co to robi? Co się z tym dzieje?

1. **być** - komputer kwantowy będzie(23), komputery kwantowe są(17), zbudowanie komputera kwantowego jest(4), komputera kwantowego są(3), komputer kwantowy jest zdolny(2),
2. **mają** - komputery kwantowe mają(12),
3. **móc** - komputery kwantowe mogą(30), komputer kwantowy mógłby(12), komputer kwantowy powstać może w ciągu kilku lat(4),
4. **iść** - komputerami kwantowymi idą więc dwoma torami(3),
5. **rozwiązują** - komputery kwantowe rozwiązują zagadnienia niedostępne naszym umysłom(9),
6. **wykonać** - komputer kwantowy może więc jednocześnie wykonać wiele rachunków równolegle(3),
7. **działać** - komputery kwantowe działają(2),

Z czym to jest powiązane? - rzeczowniki połączone z definiowanym hasłem.

1. **dostęp** - dostęp do komputera kwantowego(2),
2. **użycie** - użyciu komputera kwantowego mógłby szybko rozkładać bardzo duże liczby na iloczyny liczb pierwszych(2),
3. **teoria** - teoria komputerów kwantowych(5),
4. **praca** - prace nad komputerami kwantowymi(4),
5. **budowa** - budowę komputera kwantowego(5),
6. **przyszłość** - komputery kwantowe przyszłość informatyki(11), komputerem kwantowym jest przyszłość(2),
7. **algorytm** - algorytmy wykonywane przez komputer kwantowy są algorytmami probabilistycznymi(5),
8. **marzenie** - komputer kwantowy marzenie a realizacja(6), komputer kwantowy marzenie szpiegów(3),
9. **maszyna** - maszyna ta będzie komputerem kwantowym(2),
10. **krok** - komputer kwantowy krok naprzód komentarzy(3),
11. **obliczenie** - komputer kwantowy obliczenia(4), pojedynczy wynik obliczeń komputera kwantowego będzie niepewny(3),
12. **rozwiązania** - równoległość osiągana w komputerach kwantowych wystarczy do rozwiązania(2),
13. **układ** - komputer kwantowy układ fizyczny do opisu którego wymagana jest mechanika kwantowa zaprojektowany tak aby wynik ewolucji tego układu reprezentował rozwiązanie określonego problemu obliczeniowego(5),
14. **realizacja** - sieci optycznej dla realizacji komputerów kwantowych(5),
15. **działania** - kompletną teorię działania komputera kwantowego stworzył w połowie lat(4), podstawy działania komputera kwantowego(4),

Wnioski

- Zastosowane podejście wykorzystujące dane ilościowe dało bardzo obszerne i dobrze dopasowane do definiowanego hasła wyniki.

- W niektórych wypadkach wyszukane frazy były błędnie lub obcięte. Wiąże się to z tym, że połączenia pomiędzy kolejnymi słowami były już za mało popularne, aby algorytm dodał kolejne słowo z budowanej frazy. Czasami wynikało to z błędnych danych pobranych z Internetu.

- Ilość fraz pobranych przez robota, zależy od popularności danego hasła w Internecie. Jeżeli robot wyszuka więcej fraz w Internecie, to powstała definicja jest ostatecznie bardziej rozbudowana.

- Zbudowane w ramach pracy definicje pokazały, że dane statystyczne pozwalają na wyszukanie najważniejszych informacji związanych z danym hasłem.

Podsumowanie

- Zaprezentowany algorytm, jest dopiero jedną z pierwszych prób rozwiązania zagadnienia automatycznego generowania definicji pewnych pojęć, podjętą pod nadzorem promotora pracy dra Adriana Horzyka.

- Problem jest bardzo ciekawy, ale i kłopotliwy, z racji na swoją specyfikę. Zmiany wprowadzane w algorytmie, wymagają często wielu testów dla różnych parametrów, a to powoduje, że są one bardzo czasochłonne. Potrzeba jeszcze dużo czasu, aby dokładnie przeanalizować i wypróbować w praktyce nowe podejścia i pomysły.

- Problematyczne okazało się być wyszukiwanie pełnych zdań, które opisują dane znaczenie. Sam problem ekstrakcji pełnych zdań z Internetu, dostarcza sporo kłopotów. Do tego, nie zawsze mamy pewność, że całe zdanie opisuje definiowane pojęcie.

- Należy tutaj podkreślić fakt, że badania nad tym zagadnieniem dla języka polskiego, są bardzo potrzebne. Jeżeli sami nie będziemy rozwijać tej dziedziny, to zostaniemy w tyle za innymi krajami.

Dziękuję za uwagę!