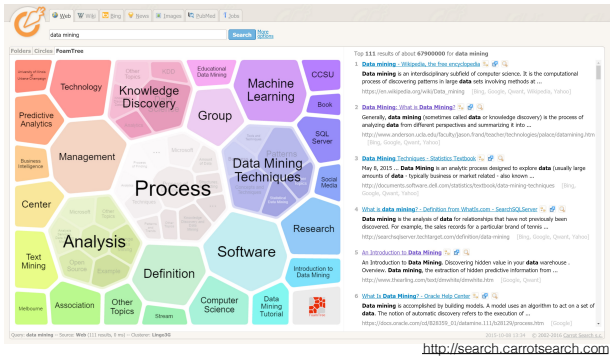
Grupowanie opisowe dużych repozytoriów danych tekstowych

Stanisław Osiński, Dawid Weiss
CARROT search

Stanisław Osiński, Dawid Weiss, Carrot Search
info@carrotsearch.com
https://carrotsearch.com



http://search.carrotsearch.com

Grupowanie opisowe

Tworzenie takich skupień, dla których da się znaleźć krótki i czytelny dla człowieka opis

Grupowanie opisowe to tworzenie takich skupień, dla których da się znaleźć krótki i czytelny dla człowieka opis. Tak jak w tradycyjnej analizie skupień, chcemy by opisy charakteryzowały zawartość grupy oraz podkreślały różnice pomiędzy grupami. W odróżnieniu od tradycyjnej analizy skupień, grupowanie opisowe może pominąć “numerycznie” istotne grupy, których algorytm nie potrafi dobrze opisać.

Przedstawiony powyżej zrzut ekranu przedstawia pogrupowanie opisowo wyniki wyszukiwania dla frazy “data mining” (do przetestowania on-line na http://search.carrotsearch.com). Większość algorytmów opisuje skupienia wykorzystując słowa lub sekwencje słów występujące w tekście (Text Mining, Social Media, Predictive Analysis).

Grupowanie opisowe (Small Data)

- **Suffix Tree Clustering**: Oren Zamir, Oren Etzioni: Web Document Clustering: A Feasibility Demonstration, SIGIR 1998.
- **Vivisimo**: obecnie część IBM Watson.
- **Carrot²**: projekt open source, zawiera naszą implementację STC oraz autorski algorytm Lingo, http://carrot2.org.

Historycznie, grupowanie opisowe stosowane było do grupowania wyników wyszukiwania (STC, Vivisimo). Nasze prace w ramach Carrot2 również koncentrowały się na małych ilościach danych.

Grupowanie opisowe (Small Data)

Jednym z podejść do grupowania opisowego jest odwrócenie typowego procesu analizy skupień: najpierw znajdujemy zbiór czytelnych i zróżnicowanych opisów (etykiet), a dopiero potem przypisujemy dokumenty do tych opisów aby utworzyć skupienia.

Istniejące algorytmy (STC, Lingo) wykorzystują frazy (częste sekwencje słów) w charakterze cech i etykiet. Selekcja etykiet może bazować na różnych kryteriach (częstości, cechy powierzchniowe, dekompozycje). Sposób przypisania dokumentów do etykiet powinien budować czytelny związek pomiędzy etykietą i dokumentem (np. dokument zawiera etykietę, słowa etykiety).

Grupowanie opisowe (Small Data)

- Zaniedbywalnie mało danych (50 - 200 kB tekstu), przetwarzanie w pamięci.
- Przetwarzanie w czasie rzeczywistym, mimo kosztownych algorytmów.
- Bezstanowa architektura, idealna do skalowania horyzontalnego.

Grupowanie opisowe w jego oryginalnym zastosowaniu wydaje się łatwe i przyjemne: mało danych, możliwość stosowania kosztownych algorytmów, architektura bezstanowa (przetwarzamy tekst wejściowy, zapominamy). Granicą stosowalności tego podejścia jest kilka-klikanaście MB tekstu.

Grupowanie opisowe (Big Data)

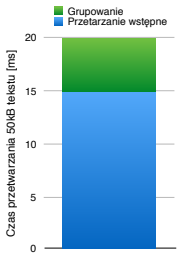
- Milion razy więcej danych (50 kB -> 50 GB), wnioski grantowe National Science Foundation.
- Zachowanie “opisowości” grup.
- Zachowanie przetwarzania w czasie rzeczywistym.
- Przetwarzanie współbieżne na jednej maszynie.

Jeden z naszych klientów (National Science Foundation) zachęcał nas i wspierał w próbach przeskalowania grupowania opisowego tak, by można było przetwarzać bazę wniosków grantowych (ok. 50 GB tekstu).

Przyjęliśmy, że chcemy obsłużyć ok. 6 rzędów więcej danych niż przy grupowaniu wyników wyszukiwania (50 kb -> 50 GB). Dla uproszczenia rezygnujemy chwilowo z przetwarzania rozproszonego, pozostajemy przy przetwarzaniu współbieżnym na jednej maszynie.

Chcieliśmy zbudować system, w którym mimo wzrostu ilości danych tekstowych, uda się w miarę możliwości utrzymać przetwarzanie w czasie rzeczywistym (czas przetwarzania liczony w sekundach lub minutach) oraz czytelne, przyjazne dla człowieka opisy skupień.

50 kB -> 50 GB



Przetwarzanie wstępne:

50 kB tekstu: 15 ms
50 GB tekstu: ~4 godziny

Rozwiązanie: składowanie wyników przetwarzania wstępnego.

Przeskalowanie poprzedniego podejścia do skali Big Data przy zachowaniu cech przetwarzania w czasie rzeczywistym napotyka na dwa podstawowe problemy.

Przetwarzanie wstępne (podział na tokeny, normalizacja, stemming, ekstrakcja i zliczanie fraz) jest relatywnie kosztowne. Przetworzenie wstępne 50 GB tekstu zajmuje klika godzin, co oddala nas od przetwarzania w czasie rzeczywistym. Jedną z możliwości jest składowanie wyników przetwarzania wstępnego (co oczywiście nie jest niczym nowym, rozwiązanie od dziesiątek lat stosowane w systemach baz danych i wyszukiwarkach pełnotekstowych). Przy takim podejściu, przetwarzanie wstępne wykonywane jest raz a sama analiza skupień, przeprowadzania dla całej kolekcji lub dla jej podzbioru, korzysta z etykiet składowanych na dysku. Warto zauważyć, że ta architektura nie jest już bezstanowa, indeks wymaga pielęgnacji (dodawanie, usuwanie dokumentów).

---

50 kB -> 50 GB

200 wycinków, ~30 grup

880k artykułów z PubMed, 1000 grup

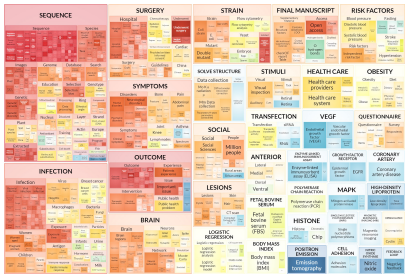
Kolejny problem wynika z charakterystyki skupień otrzymywanych przy naszym podejściu do grupowania opisowego: skupienia są raczej małe, dość precyzyjne. Pokrycie grupami 200 wyników wyszukiwania wymaga ok. 30 grup, pokrycie setek tysięcy artykułów z PubMed wymaga co najmniej 500—1000 grup (lub więcej jeśli np. wykluczymy bardzo częste etykiety). Aby zachować obecne podejście musimy ułatwić użytkownikom analizę i wybór interesujących skupień.

50 kB -> 50 GB

200 wycinków, ~30 grup

- Process (18)
- Data Mining Techniques (14)
- Data Science (14)
- Knowledge Discovery (11)
- Machine Learning (8)
- Software (8)
- Management (8)
- Business (8)
- Algorithms (8)
- Definition (7)
- Research (6)
- Tool (6)
- Text Mining (5)
- SQL Server (4)
- Data Mining Models (8)
- Companies (6)
- Group (5)
- Association (4)
- Orange Data Mining (3)
- Introduction to Data Mining (3)
- Center (3)
- KDD (3)
- Expert (3)
- Predictive Analytics (4)
- Example (4)
- Finding (5)
- Methods (5)
- Other topics (13)

880k artykułów z PubMed, 1000 grup



Jednym ze sposobów ułatwienia analizy dużej liczby skupień jest pogrupowanie małych grup w większe struktury.

Grupowanie opisowe (Big Data)

Przyjęcie pokazanych wcześniej założeń wymaga podzielenia procesu grupowania na dwie fazy: indeksowanie (które odbywa się raz) oraz analizę skupień (która korzysta z uprzednio przygotowanego indeksu).

Podczas indeksowania przetwarzamy całą kolekcję w celu wyodrębnienia cech i etykiet, przypisania dokumentów do etykiet i zapisania na dysku tych danych (w postaci indeksu odwrotnego).

Analizę skupień przeprowadzania jest dla zdefiniowanego podzbioru indeksu (w szczególności całego indeksu). Dla zadanego podzbioru następuje selekcja i grupowanie etykiet, które, wraz z przypisanymi im dokumentami, tworzą wynik analizy skupień.

Przetwarzanie wstępne

- "Naiwne", dokładne zliczanie częstości słów i fraz.
- Normalizacja interpunkcji, wielkości liter, formy gramatycznej.
- Składowanie w postaci indeksu odwrotnego (Lucene) wybranych fraz (np. DF >= 10).
- Nie udało nam się osiągnąć dobrych wyników ze zliczaniem probabilistycznym.

```
good|clinical|practice|gcp =>
  >"Good Clinical Practice (GCP)"< 1
  >"Good Clinical Practice" (GCP)< 1
  >"good clinical practice" (GCP)< 1
  >'good clinical practice' (GCP)< 1
  >Good Clinical Practice (GCP)< 303 (*)
  >Good Clinical Practice (GCP))< 1
  >Good Clinical Practice GCP< 1
  >Good Clinical Practices (GCP)< 6
  >Good Clinical Practices (GCPs)< 5
  >Good Clinical practice (GCP)< 1
  >Good clinical Practice (GCP)< 2
  >Good clinical practice (GCP)< 14
  >Good clinical practice( GCP)< 1
  >good clinical practice (GCP)< 65
  >good clinical practices (GCP)< 2
```

Ekstrakcja fraz bazuje na "naiwnym" dokładnym zliczaniu częstości słów i sekwencji słów, korzysta w razie potrzeby z dysku. Zliczaniu podlegają znormalizowane postacie cech (wielkość znaków, interpunkcja, formy gramatyczne), do składowania wybierane są frazy spełniające zadane progi częstości (np. DF >= 10). Wyniki przetwarzania wstępnego są składowane w postaci indeksu odwrotnego (Lucene). Obecna implementacja jest w stanie przetworzyć 5—15 GB tekstu na godzinę.

Eksperymentowaliśmy z probabilistycznym zliczaniem jednak bez większego sukcesu. Algorytmy te działają dobrze gdy trzeba wybrać np. 1k z 100M cech, ale gorzej gdy chcemy wybrać np. 5M cech z 100M.

Selekcja etykiet

- Dla zadanego podzbioru kolekcji, w szczególności dla całej kolekcji, wybieramy 500—5000 fraz, które staną się etykietami grup.
- Kryteria oparte na częstości występowania fraz oraz charakterystykach powierzchniowych (np. liczba słów, forma gramatyczna).
- Wykorzystanie globalnych statystyk częstotliwości.

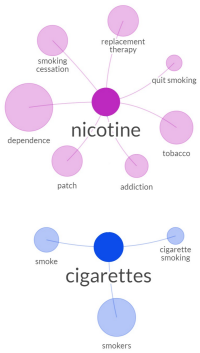
$$\frac{DF_{podzbioru} / D_{podzbioru}}{DF_{globalne} / D_{globalne}} > \text{próg}$$

Selekcja cech korzysta z danych zgromadzonych w indeksie odwrotnym i jest przeprowadzana dla zadanego podzbioru indeksu. Podobnie jak w rozwiązaniach Small Data, selekcja opiera się głównie na częstościach wystąpień (np. spełnianie zadanych progów DF) oraz charakterystyce powierzchniowej etykiet (liczba słów).

Jeśli selekcja cech przebiega dla podzbioru właściwego całej kolekcji, możemy wykorzystać globalne statystyki częstotliwości wystąpień. Często stosowaną praktyką jest selekcja cech, które, relatywnie, występują częściej w przetwarzanym podzbiorze niż globalnie w kolekcji. Tego typu możliwości nie mieliśmy w przetwarzaniu wyników wyszukiwania (ad hoc).

Grupowanie etykiet

- Algorytmy typu k-medoidów pozwalają zachować “opisowość”.
- **Affinity Propagation**: Brendan J. Frey and Delbert Dueck: Clustering by Passing Messages Between Data Points, Science 315, 972–976, 2007.



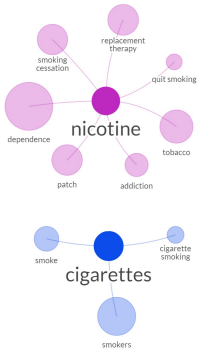
Jeśli chodzi o grupowanie etykiet, szczególnie atrakcyjne wydają się algorytmy typu k-medoidów (takie jak PAM, czy AP), które do każdego skupienia przydzielają jeden wyróżniony punkt danych, który można by uznać za pewien opis całego skupienia.

Z prawej strony pokazane są dwa przykładowe skupienia etykiet. Grupa opisana jako “nicotine” zawiera etykiety takie jak “patch”, “replacement therapy”, “tobacco”, “smokers”.

W obecnej implementacji do grupowania etykiet wykorzystujemy algorytm Affinity Propagation.

Affinity Propagation

- Wejście: macierz podobieństw.
- Wyjście: lista skupień, każde skupienie z wyróżnionym archetypem (*exemplar*).
- Iteracyjna wymiana komunikatów pomiędzy punktami danych w celu maksymalizacji sumy podobieństw do archetypów.
- Niewielkie ograniczenia co do miary podobieństwa.
- Wszystkie punkty jednocześnie rozważane jako potencjalne archetypy.



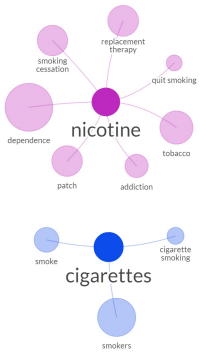
Wejściem dla algorytmu AP jest macierz podobieństw pomiędzy grupowanymi punktami (w naszym przypadku — etykietami). Na wyjściu otrzymujemy listę skupień z przypisanymi do nich wyróżnionymi etykietami-archetypami (exemplar).

Algorytm AP polega na iteracyjnej wymianie komunikatów pomiędzy punktami w celu znalezienia rozwiązania, które maksymalizuje sumę podobieństw punktów do przypisanych im archetypów.

Złożoność każdej iteracji zależy liniowo od liczby elementów w macierzy podobieństw (czyli w najgorszym wypadku od kwadratu liczby etykiet). Ponieważ na etapie selekcji cech wybraliśmy podzbiór 500—5000 cech, czas potrzebny na wykonanie algorytmu AP pozostaje w sensownych granicach.

Affinity Propagation

- Wejście: macierz podobieństw.
- Wyjście: lista skupień, każde skupienie z wyróżnionym archetypem (*exemplar*).
- Iteracyjna wymiana komunikatów pomiędzy punktami danych w celu maksymalizacji sumy podobieństw do archetypów.
- Niewielkie ograniczenia co do miary podobieństwa.
- Wszystkie punkty jednocześnie rozważane jako potencjalne archetypy.



Affinity Propagation ma kilka ciekawych cech:

- Nie nakłada żadnych ograniczeń na macierz podobieństwa, podobieństwa mogą być nawet asymetryczne.
- Wszystkie punkty jednocześnie są rozważane jako potencjalne archetypy.
- Liczba skupień nie jest zadawana wprost, wynika z wartości “preferencji” przypisanej każdemu punktowi (preferencja określa jak silnie dany punkt powinien być preferowany jako archetyp w ostatecznym wyniku).

Miara podobieństwa

- Liczba współwystąpień par etykiet w zadanym oknie.

	nicotine	cigarettes	cigarette smoking
nicotine		206	38
cigarettes	206		532
cigarette smoking	38	532	
	1088	1128	534

- Sposób ważenia (Dice, Jaccard, Simpson, …) wpływa na typ relacji cech wewnątrz skupień.

	nicotine	cigarettes	cigarette smoking
nicotine		0.19	0.07
cigarettes	0.19		0.99
cigarette smoking	0.07	0.99	

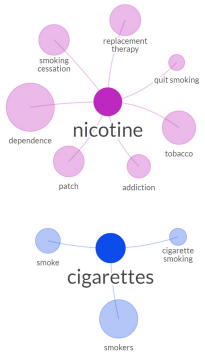
Miara podobieństwa pomiędzy etykietami zależy od liczby współwystąpień w zadanym oknie (np. w odległości co najwyżej 16 tokenów). Sposób ważenia liczby współwystąpień (Dice, Jaccard, Simpson) wpływa na charakterystykę skupień (liczba, wielkość) i na charakter relacji pomiędzy etykietami wewnątrz skupień.

W celu uzyskania bardziej “semantycznych” grup etykiet wykorzystuje się często współwystąpienia drugiego rzędu (odległości pomiędzy wektorami reprezentującymi współwystąpienia pierwszego rzędu). Planujemy przetestować również i to podejście, chociaż proste współwystąpienia działają zaskakująco dobrze, być może to kwestia samego algorytmu grupującego.

Affinity Propagation, rozszerzenia

M. Leone, Sumedha, M. Weigt: Clustering by soft-constraint affinity propagation: applications to gene-expression data. Bioinformatics, 2007.

Y. Fujiwara, G. Irie, T. Kitahara: Fast algorithm for affinity propagation. IJCAI'11, 2011.



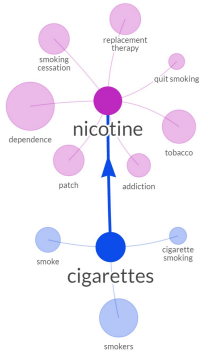
W oryginalnym sformułowaniu AP wymaga, by archetyp należał do skupienia, które definiuje (innymi słowy: by archetyp był archetypem dla samego siebie). W konsekwencji otrzymujemy listę izolowanych nienakładających się skupień etykiet.

Ciekawym rozszerzeniem oryginalnego algorytmu AP jest Soft-Constraint AP, gdzie to wymaganie jest poluźnione — punkt, który jest archetypem dla innych punktów może wybrać jeszcze inny punkt (a nie siebie) jako archetyp. W ten sposób zezwalamy na powstawanie połączeń pomiędzy niektórymi skupieniami, otrzymując pewną liczbę izolowanych grafów etykiet.

Affinity Propagation, rozszerzenia

M. Leone, Sumedha, M. Weigt: Clustering by soft-constraint affinity propagation: applications to gene-expression data. Bioinformatics, 2007.

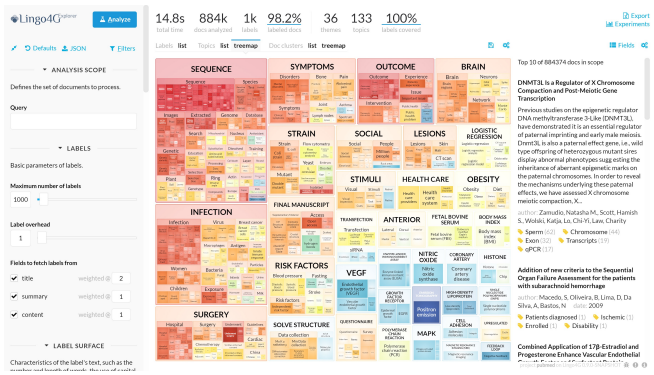
Y. Fujiwara, G. Irie, T. Kitahara: Fast algorithm for affinity propagation. IJCAI'11, 2011.



Zastosowanie Soft-Constraint AP dodaje zatem kolejny, wyższy, poziom struktury:

- Pierwszy, najniższy, poziom stanowią skupienia dokumentów definiowane przez same etykiety.
- Drugi poziom stanowią skupienia etykiet wokół archetypów ("nicotine" oraz "cigarettes" w przykładzie po prawej stronie).
- Trzeci, najwyższy, poziom stanowią spójne grafy skupień etykiet (wszystkie etykiety widoczne w przykładzie).

Inny bardzo ciekawy pomysł, specyficzny dla AP, to możliwość eliminacji części elementów macierzy podobieństwa bez wpływu na ostateczny wynik grupowania. Efektywność tego podejścia mocno zależy od charakteru danych. Przy naszym zastosowaniu (grupowanie na podstawie współwystąpień) często udaje się usunąć nawet 30-50% elementów.



http://get.carrotsearch.com/lingo4g/latest/doc

Wszystkie rozwiązania opisane w tej prezentacji wchodzą w skład tworzonego przez nas systemu grupowania opisowego. System ten jest obecnie testowany przez zamkniętą grupę użytkowników testowych.

Dla zainteresowanych, szczegóły techniczne rozwiązania dostępne są w dokumentacji: http://get.carrotsearch.com/lingo4g/latest/doc.

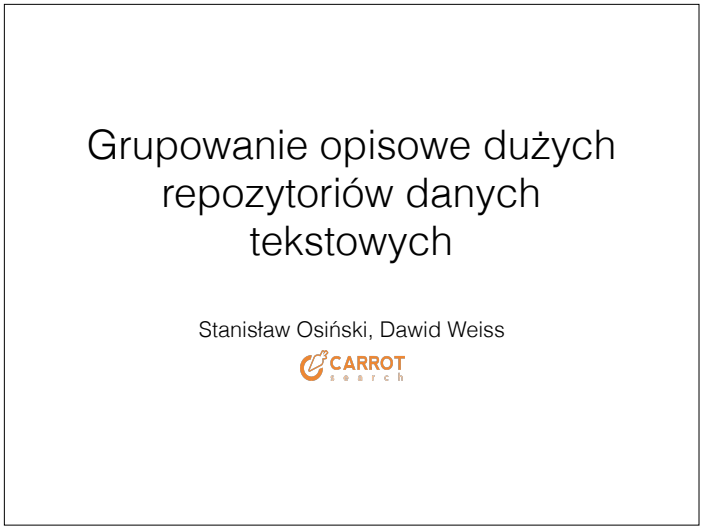
Przyszłość

- Mapy 2D (t-SNE, t-Distributed Stochastic Neighbor Embedding).
- “Semantyczne” miary podobieństwa, Random Indexing.
- Równoczesne grupowanie etykiet i dokumentów (co-clustering).

W przyszłości chcielibyśmy poeksperymentować z prezentacją etykiet w postaci dwuwymiarowej mapy, na której powiązane etykiety byłyby blisko siebie. Dość modnym algorytmem jest tu obecnie t-SNE, zrzut ekranu pokazuje prototyp, który jakiś czas temu zbudowaliśmy.

Chcielibyśmy też potestować “semantyczne” miary podobieństwa (np. oparte na współwystąpieniach drugiego rzędu). Implementacja w tej skali będzie pewnie wymagała zastosowania specjalnych technik, np. Random Indexing.

Ciekawe byłoby też równoczesne grupowanie dokumentów i etykiet.



Stanisław Osiński, Dawid Weiss, Carrot Search
info@carrotsearch.com
https://carrotsearch.com