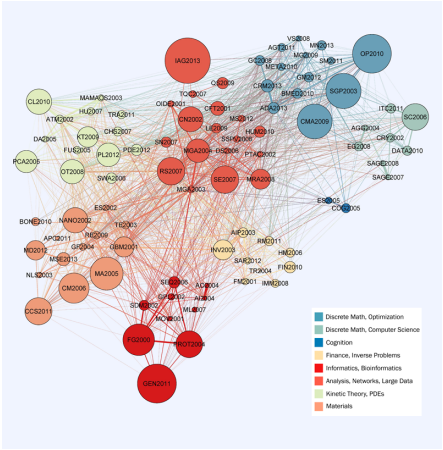
Zarówno klasyfikacja wieloetykietowa jak i znajdowanie skupisk w sieciach społecznych są zjawiskami z którymi spotykamy się już na codzień. Dzięki klasyfikacji możemy wyszukiwać muzykę w Spotify po gatunku, albo filmy na Youtube po tematyce. Klasyfikacja wieloetykietowa polega na uczeniu - tzw. klasyfikatora - relacji między cechami obiektów a etykietami. Klasyfikacja wieloetykietowa zakłada możliwość przyporządkowania więcej niż jednej etykiety w przeciwieństwie do klasycznego problemu klasyfikacji, który pozwalał przyporządkować tylko jedną klasę. Najczęscie klasyfikacji wieloetykietowej jednak dokonuje się redukując problem wieloetykietowy do problemu przydzielenia poprawnej klasy.

Skupiska w sieciach Znajdowanie skupisk to próba znalezienia grup przedmiotów, filmów, książek, produktów itd., które są do siebie bardziej podobne ponieważ są połączone większą liczbą relacji między sobą wewnątrz skupiska niż z innymi obiektami. Najprostszym przykładem takich relacji jest fakt, że dwie osoby kupiły ten sam produkt, przeczytały te samą książkę itd. Bardziej rozbudowanym - np. fakt, że dwie osoby tak samo oceniły film, który widziały. Relacja może mieć przyporządkowaną wagę - miarę ważności w stosunku do innych relacji, np. taką wagą może być liczba osób, które wystawiły recenzje o tej samej ocenie, albo prościej - liczba osób, która kupiła te dwa produkty. Na przykładowym Rysunku 1 znajdowanie skupisk zostało zastosowane na grafie reprezentującym powiązania między programami i warsztatami matematycznymi, zaś krawędzie reprezentują współuczestnictwo naukowców w danym programie/warsztacie. Waga krawędzi, ukazana przez grubość, to procent uczestników tych wydarzeń, którzy wzięli udział w obu. [1]

Rysunek 1: Znalezione skupiska dzielą zbiór programów wsparcia nauki na dziedziny - umożliwiając dokładniejsza klasyfikację wewnątrz dziedzin



Skupiska a klasyfikacja Dziś spotykamy się z koniecznością klasyfikacji co raz większych zbiorów danych. Im zaś większe mamy zbiory, tym więcej mamy różnych kombinacji etykiet do przydzielenia, choć wcale nie jest tak, że na każdą kombinację przypada po tyle samo obiektów. Najczęściej jest zupełnie inaczej - na pewne kombinacje przypada ogromna większość obiektów niż na inne. Łatwo można sobie wyobrazić, że klasyfikując filmy muzyczne na Youtube, więcej dostaniemy coverów znanych gwiazd muzyki pop, niż wykonań utworów klasycznych.

Metody, którymi dysponujemy w klasyfikacji, polegają na wyszkukiwaniu istotnych różnic między cechami obiektów, np. w przypadku klasyfikacji coverów muzyki, takimi cechami mogłoby być tempo, dynamika. Różnicując te obiekty na bazie ich cech możemy przyporządkować rodzaj muzyki czy instrumenty na których jest wykonywana, pod warunkiem zachowania proporcji między dokładnością z jaką metoda potrafi odróżniać od siebie dwa wykonania, a skalą różnic. Naturalne jest, że w sytuacji gdy będziemy dysponować tylko wykonaniami z jednego gatunku muzycznego łatwiej będzie nauczyć metodę bardziej wyrafinowanych różnic między nimi - co spowoduje lepsze otagowanie np. konkretnym wykonawcą, niż w sytuacji gdy tych samych różnic będziemy próbowali wyuczyć się w zbiorze, który jednoczesnie składa się z muzyki pop i muzyki klasycznej.

Aby uzyskać możliwość klasyfikacji w spójnych podgrupach, chcemy wykorzystać metody znajdowania skupisk z sieci społecznych do dzielenia wyjściowej przestrzeni etykiet. Do tej pory stosowano raczej metody dzielenia na podstawie analizy bliskości cech obiektów, podział zaś był dokonywany na rozłączne skupiska. Uważamy, że wydajność tych metod można poprawić opierając się na eksploracji relacji między etykietami i na ich bazie budowania podziału przestrzeni klasyfikacji. Dlatego niniejszym projekcie chcemy zweryfikować czy i na ile można poprawić wydajność klasyfikacji wieloetykietowej poprzez podział zbioru etykiet na podzbiory, które mogą na siebie zachodzić - a w sytuacji zachodzenia, dobrać odpopowiednią metodę podejmowania decyzji o tym czy etykietę przyporządkować czy nie. Planujemy podjąć ten temat zarówno z perspektywy płaskiej jak i hierarchicznej klasyfikacji. Nasze wstepne wyniki dają zadowalające rezultaty[2].

[1] http://helper.ipam.ucla.edu/articles/ipamconnections.aspx
[2] https://arxiv.org/abs/1606.02346