

# Studium przypadku: struktura wiedzy w badaniach rynku stali

Lee Hyunjong, Sohn Il

W życiu codziennym zajmujemy się danymi na większą i na mniejszą skalę. Z perspektywy państw ważne są wartości produktu krajowego i wzrost ekonomiczny, a dla firm najistotniejsze są wyniki sprzedaży. Pracownicy naukowcy analizują dane i wyniki badań, a studenci używają swoich ocen jako danych, które pozwalają im wstępować na uniwersytety. Z tego punktu widzenia nie ma nikogo, kto by nie miał do czynienia z żadnymi danymi.

Jednak wzrokowe znalezienie informacji w danych nie jest łatwe. Szczególnie dotyczy to obecnej ery Big Data, w której różnego rodzaju dane gromadzi się bardzo szybko i na ogromną skalę. Dane w jakiś sposób pokazują rzeczywistość, ale nie wystarczy się im przyjrzeć, aby je zrozumieć. Aby uzyskać zawarte w nich informacje, trzeba te dane przeanalizować.

Naszym zamiarem było udostępnienie pracownikom naukowym przydatnych informacji, uzyskanych poprzez analizę rzeczywistych danych. Podjęliśmy próbę analizy „struktury wiedzy” w badaniach rynku stali, a wyniki tej pracy naukowej zostały opublikowane na łamach „Steel Research International”. W ten sposób chcieliśmy usystematyzować dotychczasowe badania branży stalowej i stworzyć przewodnik przydatny w rozwijaniu tych badań. Rozwój badań w przemyśle stalowym następuje szybko, zarówno w aspekcie akademickim, jak i praktycznym. Zadaliśmy sobie pytanie: jakie cechy wyróżniają ten proces rozwoju? Osoby biegłe w interpretacji danych lub prowadzące badania w określonej dziedzinie potrafią je wychwycić już po zapoznaniu się z jakimiś dokumentami badawczymi albo liczbami. Jednak dla przeciętnego człowieka są to tylko listy liczb i podobnych elementów. Dlatego właśnie potrzebna nam była metoda wyszukiwania pewnych cech lub struktur z samych danych, czyli z samej rzeczywistości. I to właśnie nazywamy „analizą sieci Big Data”. Poniżej szczegółowo omawiamy sposób zbierania i analizowania danych, które zostały opublikowane w „Steel Research International” [1] – numer ze stycznia 2015 r. Dzięki temu można zrozumieć strukturę wiedzy w przemyśle stalowym.

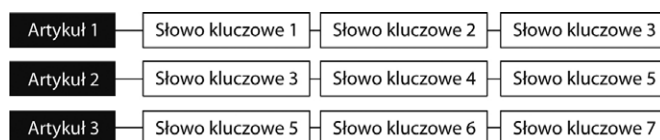
**Etap 1: Zbieranie danych.** Zbieranie informacji bibliograficznych z artykułów opublikowanych w periodyku „Steel Research International”. Są to artykuły akademickie, recenzje, analizy przypadków, korekty, recenzje książek, wytyczne redakcyjne itd. Z tego zbioru pozostawiliśmy tylko artykuły akademickie, recenzje i analizy przypadków, ponieważ inne nie mają bezpośredniego wpływu na strukturę wiedzy w badaniach rynku stali.

**Etap 2: Wyodrębnianie danych.** Wyodrębnianie słów kluczowych z zebranych artykułów. Te słowa kluczowe wskazują

tematykę artykułów, więc można je uzyskać z list słów kluczowych, abstraktów i tytułów artykułów.

**Etap 3: Oczyszczanie danych.** Różni autorzy mogą różnie zapisywać słowa kluczowe, więc trzeba wykonać ich standaryzację. W oczyszczaniu zbioru słów kluczowych pomogli nam również eksperci z branży stalowej.

**Etap 4: Porządkowanie danych.** Aby sieć można było analizować pod kątem struktury wiedzy, dane muszą zostać odpowiednio uporządkowane. W tym przypadku utworzyliśmy sieć dwumodalną z artykułami i słowami kluczowymi, a później, do celów analizy rocznikowej, dodaliśmy lata publikacji (patrz rys. 1).



Rys. 1. Sieć dwumodalna (artykuł - słowo kluczowe)

**Etap 5: Przekształcanie danych w sieć jednomodalną.** Trudno wykryć strukturę słów kluczowych na podstawie kierunkowych relacji w sieci dwumodalnej, czyli relacji między artykułami i słowami kluczowymi (patrz rys. 2).

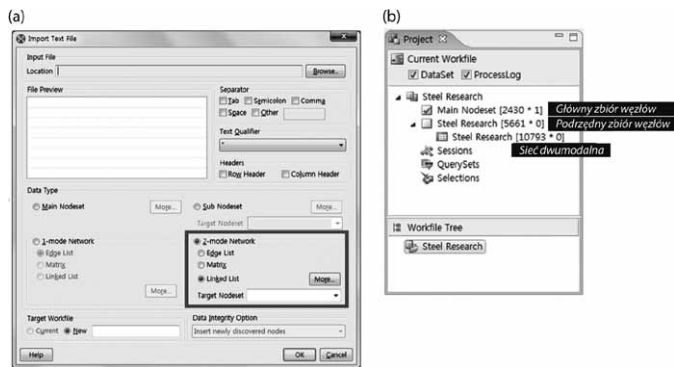


Rys. 2. Sieć jednomodalna (słowo kluczowe - słowo kluczowe)

**Etap 6: Przeprowadzenie analizy danych sieciowych.** Etap ten obejmuje analizę centralności i spójnych podgrup.

**Etap 6.1:** Wysoki wskaźnik centralności uzyskany w analizie centralności pewnych słów kluczowych oznacza, że wiele słów kluczowych występuje jednocześnie w tych samych artykułach. Wysoki wskaźnik centralności bliskości oznacza krótką odległość geodezyjną od innych słów kluczowych w sieci, ujawnia więc pozycję, która najszybciej może wpływać na całość badań rynku stali. Słowo kluczowe z wysokim wskaźnikiem centralności pośredniczenia łączy obszary badawcze w branży stalowej. Te słowa kluczowe prowadzą do zbieżności badań.

**Etap 6.2:** Przeprowadzamy analizę spójnych podgrup, aby zrozumieć podobszary w badaniach stali. Zbiór słów



**Rys. 3.** Ekran wyświetlany w trakcie importowania danych: (a) importowanie danych sieciowych; (b) plik roboczy po zaimportowaniu danych do programu NetMiner

kluczowych występujących jednocześnie w wielu artykułach można uznać za jeden z obszarów badań.

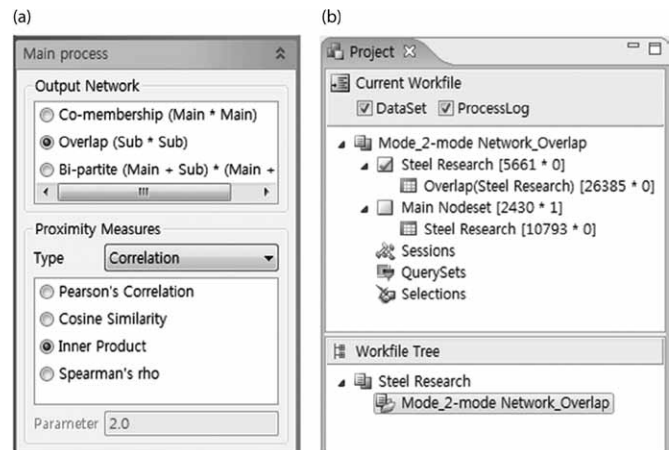
**Etap 7:** Przeprowadzamy dodatkowe analizy danych. Wykonujemy dodatkowe analizy badawcze zgodnie z celem badania. W tym przypadku utworzyliśmy mapy na podstawie wyników badań, aby na bazie spójnych podgrup zrozumieć kierunek rozwoju badań stali.

Następnie została przeprowadzona analiza, której celem było zrozumienie struktury wiedzy w badaniach rynku stali. Teraz zajmiemy się czwartym etapem, czyli porządkowaniem danych. Analiza sieci została przeprowadzona w programie NetMiner.

**Etap 4:** Organizacja zebranych danych zaczyna się od utworzenia sieci dwumodalnej, takiej jak na rysunku 3 a. Importujemy zebrane dane do programu NetMiner (*File > Import > Text File lub Excel File*).

Jeśli dane są importowane z Excela, nie trzeba konfigurować dodatkowego separatora, ale w przypadku danych z pliku tekstowego dzielimy kolumny danych za pomocą opcji Separator. Ponadto w opcjach importowania zaznaczamy *Linked List* w ramce *2-mode Network*, ponieważ zebrane dane mają postać prostej listy relacji w sieci dwumodalnej. Po zaimportowaniu danych w bieżącym pliku NetMinera zostanie utworzony główny zbiór węzłów, podrzędny zbiór węzłów i sieci dwumodalne. W drzewie pliku roboczego widać, że utworzona nazwa pliku roboczego zgadza się z nazwą importowanego pliku (patrz rys. 3 b).

**Etap 5:** Dane importowane do NetMinera to sieć dwumodalna złożona z par artykuł – słowo kluczowe. Aby można było analizować strukturę wiedzy w badanej dziedzinie, trzeba przekształcić tę sieć w sieć jednomodalną zbudowaną z par słowo kluczowe – słowo kluczowe. Sieć dwumodalną możemy przekształcić w sieć jednomodalną, wydając polecenie *Transform > Mode > 2-mode Network*. Następnie w ramce *Output Network* w okienku *Main process* zaznaczamy *Overlap (Sub\*Sub)*, a w ramce *Proximity Measures* wybieramy kolejno *Type > Correlation > Inner Product*



**Rys. 4.** Przekształcanie sieci jednomodalnej w dwumodalną: (a) przekształcanie sieci; (b) wygenerowany plik roboczy

(patrz rys. 4 a). Gdy to zrobimy, klikamy *Run Process* i pojawia się kontekstowe menu z pytaniem, czy utworzyć plik roboczy. Jeśli klikniemy *Yes*, wówczas istniejące dane będą zachowane i zostanie wygenerowany podrzędny plik roboczy, taki jak na rysunku 4 b. W programie NetMiner liczby znajdujące się po nazwach zbioru węzłów i sieci oznaczają odpowiednio liczbę węzłów i atrybutów węzłów oraz liczbę łączy i atrybutów łączy. Na przykład *Steel Research [5661\*0]* w nowo wygenerowanej sieci jednomodalnej oznacza, że całkowita liczba węzłów to 5661 i nie ma żadnych atrybutów węzłów. Natomiast *Overlap (Steel Research) [26385\*0]* w sieci jednomodalnej oznacza, że całkowita liczba łączy to 26385 i nie ma atrybutów łączy.

Wygenerowana macierz współwystępowania słów kluczowych wygląda tak jak w tabeli 1. Można ją zobaczyć w jednomodalnej sieci *Overlap (Steel Research)*. Każda komórka tej macierzy zawiera informację o częstotliwości współwystępowania słowa kluczowego *i* oraz słowa kluczowego *j* w okresie objętym badaniem (lata 1990–2013), natomiast na przekątnej widzimy częstość słowa kluczowego *i* w tym samym okresie. W szczególności w macierzy współwystępowania widzimy, że słowo *microstructure* (mikrostruktura) pojawiło się w ciągu 24 lat jako słowo kluczowe w 95 pracach naukowych, a w 4 z nich wspólnie z *duplex stainless steel* (stal nierdzewna typu duplex).

**Tabela 1.** Macierz występowania par słów kluczowych

	Nitrogen	Annealing	Duplex Stainless Steel	Microstructure	Direct Reduction
Nitrogen	28	1	1	1	
Annealing	1	10	1		
Duplex Stainless Steel	1	1	11	4	
Microstructure	1		4	95	
Direct reduction					5

**Tabela 2.** Częstość występowania słów kluczowych w latach 1990–2013

Słowa kluczowe	Liczba w całym okresie	Słowa kluczowe	1990–1994	Słowa kluczowe	1995–1999	Słowa kluczowe	2000–2004	Słowa kluczowe	2005–2009	Słowa kluczowe
Microstructure	95	Microstructure	14	Kinetics	15	Transformation induced plasticity	17	Finite element method	37	Microstructure
Finite element method	79	Heat transfer	10	Liquid iron	11	Continuous casting	14	Microstructure	29	Transformation induced plasticity
Transformation induced plasticity	72	Kinetics	10	Heat transfer	8	Mechanical properties	14	Continuous casting	22	Twinning induced plasticity
Mechanical properties	64	Mass transfer	9	Mechanical properties	8	Microstructure	12	Inclusion	20	Finite element method
Continuous casting	62	Nitrogen	9	Nitrogen	8	Finite element method	10	Slag	19	Blast furnace
Slag	54	Hot rolling	8	Finite element method	7	Slag	10	Transformation induced plasticity	19	Continuous casting
Blast furnace	50	Liquid iron	8	High temperature	6	Aluminum	7	Blast furnace	18	Mechanical properties
Kinetics	42	Thermodynamic	8	Iron	6	Blast furnace	7	Mechanical properties	16	Precipitation
Stainless steel	41	Mechanical properties	7	Liquid steel	6	Low carbon steel	7	Stainless steel	14	Slag
Heat transfer	39	Niobum	7	Low carbon steel	6	Solidification	7	High strength steel	11	Phase transformation

Z macierzy współwystępowania możemy więc odczytać częstość występowania słów kluczowych w badanym okresie. Co więcej, dzieląc tę macierz na lata, można ustalić, jakie badania dominowały w każdym roku (patrz tab. 2).

Dodatkowo możemy również zbadać częstość występowania par słów kluczowych, tak jak w tabeli 3. Dzięki temu mamy możliwość znajdowania słów kluczowych towarzyszących wybranym słowom kluczowym, na bazie których prowadzone są badania. Po wykonaniu wizualizacji wyników możemy nawigować po słowach kluczowych, a także eksplorować inne słowa kluczowe.

**Etap 6:** Na podstawie wyniku przekształcenia w sieć jednodalną można zobaczyć, które słowa kluczowe i podobszary badań są uważane za istotne w badaniach przemysłu stalowego. W tym celu przeprowadzamy analizę centralności i spójnych podgrup.

**Etap 6.1:** Aby wykonać analizę centralności, wybieramy pożądaną metodę analizy z menu *Analyze > Centrality*. W tym badaniu została przeprowadzona analiza

centralności stopnia (*Degree*), pośredniczenia dla węzłów (*Betweenness > Node*) i bliskości (*Closeness*). Chcąc obliczyć centralność stopnia, w okienku *Main process* wybieramy metodę wyznaczenia tego wskaźnika. Podstawą obliczania może być liczba wszystkich łączy do innych węzłów albo suma wag łączy. Na ogół przyjmujemy, że istnienie łącza jest ważniejsze niż jego waga i w oparciu o to założenie mierzymy centralność stopnia, ale czasami przydatna jest również druga opcja. W analizie centralności bliskości istotna jest osiągalność całej sieci, więc w okienku *Pre-process* jest domyślnie zaznaczona opcja *Dichotomize*. W okienku *Main process* trzeba również ustawić opcję w ramce *Unreachable Handling* (jak traktować węzły nieosiągalne). Najczęściej zaznaczamy *Ignore Unreachable* (ignoruj nieosiągalne). W analizie centralności pośredniczenia interesują nas również węzły i ich położenie, więc nie musimy uwzględniać wagi. Aby nie uwzględniać wag łączy, w okienku *Pre-process* domyślnie jest zaznaczone pole wyboru *Dichotomize* (patrz rys. 5).

Wynik analizy centralności w programie NetMiner przedstawiono na zakładkach [R] *Main*, [T] *Centrality Vector*, [M] *Spring* i [M] *Concentric*. [R] oznacza tu raport, [T] – tabelę, a [M] – mapę. Tak więc rezultatem analizy centralności jest jeden raport, jedna tabela i dwie mapy. Zakładka [R] *Main* zawiera informacje o procesie i podsumowanie wyników. Rozkład wskaźnika centralności i wskaźnik centralizacji sieci są pokazane w punkcie *Output Summary*. Z kolei wskaźniki centralności każdego z węzłów można zobaczyć na zakładce [T] *Centrality Vector*. Mapa z wizualizacją, wygenerowana za pomocą

**Tabela 3.** Częstość występowania par słów kluczowych

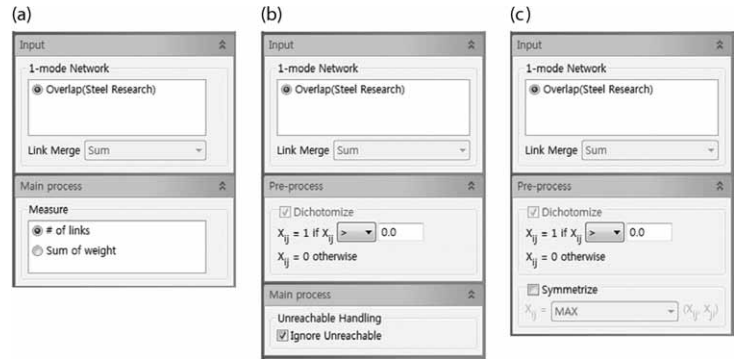
Pary słów kluczowych	Częstość	Pary słów kluczowych	Częstość
Mechanical properties – microstructure	21	Kinetic – reduction	7
Transformation induced plasticity – twinning induced plasticity	12	Viscosity – slag	7
Nitrogen – carbon	10	Al2O3 – MgO	6
Transformation induced plasticity – retained austenite	9	Blast furnace – ironmaking	6
Microstructure – heat treatment	8	Desulfurization – hot metal	6



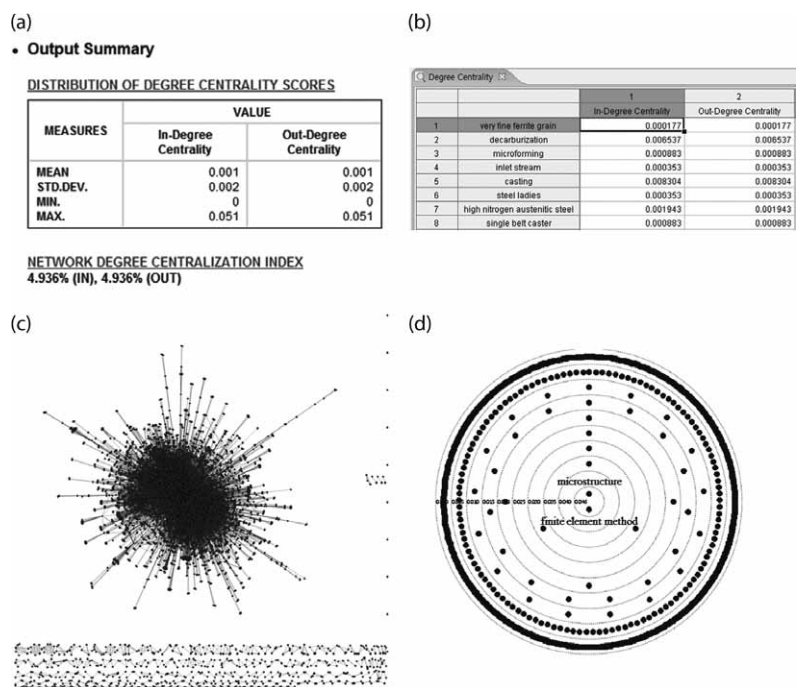
algorytmu sprężynowego, znajduje się na zakładce [M] *Spring*. Na zakładce *Display* (w dolnej części obszaru sterowania procesem) można zmienić styl mapy. Na mapie [M] *Concentric* węzły z wyższym wskaźnikiem centralności znajdują się bliżej środka, więc z łatwością można wyszukać węzeł z najwyższym wskaźnikiem. W tym przykładzie pokazano analizę centralności stopnia. W macierzy współwystępowania nie uwzględnia się kierunku, więc wskaźniki centralności stopnia, wchodzący i wychodzący, mają jednakowe wartości.

Uporządkowane wyniki analizy centralności w badaniach branży stalowej pokazano w tabeli 4. Słowa kluczowe *microstructure* i *finite element method* (metoda elementów skończonych) mają wysokie wskaźniki centralności stopnia i pośredniczenia, więc oba są uwzględniane w badaniach i łączą podobszary badań, co sprzyja badaniom zbieżności słów kluczowych. Poza tym słowa kluczowe *microstructure*, *mechanical properties* (właściwości mechaniczne) i *continuous casting* (odlewanie ciągle) znajdują się na miejscach, które sprzyjają rozpowszechnianiu się tych słów w dalszych badaniach. Dlatego też pracownicy nauki, którzy dopiero zaczynają zajmować się badaniami rynku stali, powinni wziąć pod uwagę te słowa kluczowe.

**Etap 6.2:** Przed analizą spójnych podgrup okazało się, że 94,2% wszystkich słów kluczowych w latach 1993–2013 pojawia się mniej niż pięć razy, uznając więc, że mają one stosunkowo niewielki wpływ na badania, postanowiliśmy je wykluczyć z analizy. W tym celu trzeba najpierw do głównego zbioru węzłów (*Steel Research [5661\*0]*) dodać częstotliwość jako atrybut węzła. Następnie, na pasku narzędzi NetMinera, klikamy przycisk *Query Composer*,



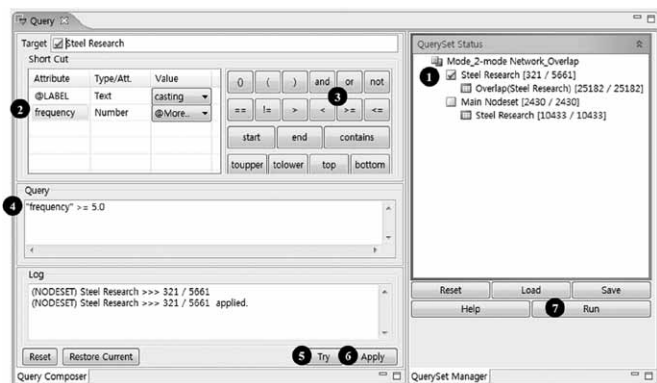
Rys. 5. Obszar sterowania procesem w trakcie analizy centralności: (a) centralność stopnia; (b) centralność bliskości; (c) centralność pośredniczenia



Rys. 6. Wyniki centralności stopnia: (a) [R] Main Report; (b) [T] Degree Centrality Vector; (c) [M] Spring; (d) [M] Concentric

Tabela 4. Analiza centralności słów kluczowych

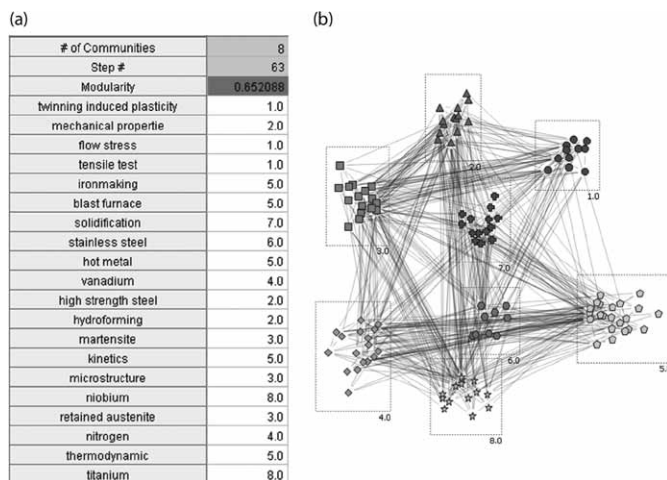
Słowo kluczowe	Centralność stopnia	Słowo kluczowe	Centralność bliskości	Słowo kluczowe	Centralność pośredniczenia
Microstructure	5,05	Microstructure	34,40	Microstructure	9,53
Finite element method	4,72	Mechanical properties	32,57	Finite element method	7,71
Continuous casting	3,87	Continuous casting	32,41	Continuous casting	6,43
Transformation induced plasticity	3,50	Slag	31,87	Slag	5,35
Mechanical properties	3,30	Kinetics	31,74	Mechanical properties	5,05
Slag	3,29	Solidification	31,55	Transformation induced plasticity	4,67
Blast furnace	2,90	Stainless steel	31,52	Kinetics	4,66
Kinetics	2,44	Transformation induced plasticity	31,38	Blast furnace	4,03
Stainless steel	2,33	Phase transformation	31,35	Stainless steel	3,61
Precipitation	2,07	Precipitation	31,33	Solidification	2,85



Rys. 7. Kompozytor zapytań

aby w obszarze edycji danych pojawiło się to, co widać na rysunku 6. W tym przykładzie, po wybraniu zbioru węzłów lub sieci w okienku *QuerySet Status*, wpisujemy zapytanie (ramka *Query*) i klikamy kolejno *Try* > *Apply* > *Run*, aby wyodrębnić słowa kluczowe, które występują więcej niż pięć razy. Tym sposobem spośród 5661 słów kluczowych do analizy spójnych podgrup wyodrębniliśmy 321, które pojawiają się częściej niż pięciokrotnie (rysunek 7).

Następnym krokiem było przeprowadzenie analizy wspólnoty. W ten sposób powstało więcej wewnętrznych (w grupach) niż zewnętrznych (między grupami) połączeń słów kluczowych, co pozwoliło skupić się na analizie podobszarów badań. Aby wykonać analizę wspólnoty, wybieramy kolejno *Analyze* > *Cohesion* > *Community* > *Modularity*. W tym przykładzie, w okienku *Main Process* na obszarze sterowania procesem, należy w ramce *Algorithms* wybrać algorytm. Najpopularniejszym algorytmem jest CNM<sup>1</sup>, który gwarantuje przyzwoite rezultaty, chociaż ze wszystkich dostępnych opcji działa najwolniej. Gdy chcemy podzielić sieć na określoną liczbę wspólnot, niezależnie od wskaźnika modułowości, korzystamy z opcji *Include*



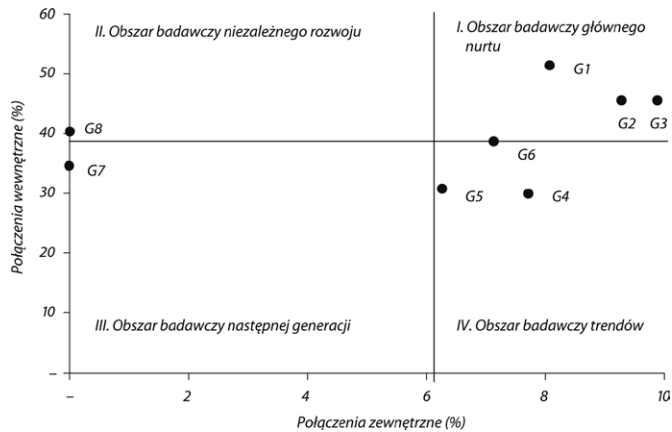
Rys. 8. Wyniki analizy spójnej podgrupy: (a) [T] *Community Partition*; (b) [M] *Clustered*

*Nonoptimal Output*. Wyniki analizy wspólnoty pojawiają się na zakładkach [R] *Main*, [T] *Community Partition* i [M] *Clustered*. *Raport* zawiera wskaźnik najlepszej modułowości, natomiast z tabeli na zakładce [T] *Community Partition* możemy odczytać informacje o przynależności węzłów do podgrup. Na zakładce [M] *Clustered* znajduje się mapa klastrowania spójnych podgrup, wygenerowana na podstawie wskaźnika modułowości. W tym przykładzie mapy klastrowania każdej spójnej podgrupy mogą mieć różne kształty i kolory, które ustawiamy w obszarze sterowania procesem, na zakładce *Display* (okienko *Node Style*, ramka *Node Attribute Styling*) (patrz rys. 8).

W wyniku analizy spójnych podgrup w badaniach stali wyodrębniono w sumie osiem podgrup (patrz tab. 5). Liczbę słów kluczowych, liczbę łącz i połączenia wewnętrzne (gęstość) można znaleźć za pomocą polecenia *Analyze* > *Properties* > *Group*. Informacje o połączeniach zewnętrznych uzyskujemy w wyniku dodatkowej analizy badawczej.

Tabela 5. Podobszary badań stali

Numer	Spójne podgrupy		Liczba słów kluczowych	Liczba łącz	Połączenia wewnętrzne (%)	Połączenia zewnętrzne (%)	Kategoria badawcza
	Słowa kluczowe należące do podgrup						
G1	Austenitic steel, ferritic stainless steel, flow stress, grain size, hot deformation, martensitic transformation itd.		11	56	50,90	8,15	I
G2	Austenite, austenitic stainless steel, ductility, fatigue, finite element method, high strength steel, hot stamping itd.		14	82	45,10	9,36	I
G3	Annealing, cold rolling, deformation, dual phase steel, duplex stainless steel, electron backscattered diffraction, formability itd.		16	108	45,00	9,90	I
G4	Activity, alloy, aluminum, copper, high speed steel, iron, liquid iron itd.		17	80	29,40	7,76	IV
G5	Basic oxygen furnaces, basic oxygen furnaces slag, blast furnace, carburization, computational fluid dynamic, decarburization, dephosphorization itd.		22	138	29,90	6,32	IV
G6	Continuous casting, high temperature, molten steel, mold, optimization, physical modeling, surface tension itd.		9	28	38,90	7,20	I
G7	Casting, characterization, composition, fluid flow, heat transfer, inclusion, ladle itd.		14	62	34,10	0,06	III
G8	Carbide, carbon, corrosion, creep, hardness, microalloy, microalloyed steel itd.		14	72	39,60	0,08	II



Rys. 9. Mapowanie obszarów podgrup w badaniach stali

**Etap 7:** Aby lepiej zrozumieć podobszary badań przemysłu stalowego, kategorie badawcze podzielono na połączenia wewnętrzne i zewnętrzne (patrz rys. 9). Podgrupy w podobszarach mających wysoki wskaźnik zarówno połączeń wewnętrznych, jak i zewnętrznych, to obszar badawczy głównego nurtu (ang. *main stream research area*), który zasługuje na uwagę jako rozwijająca się dziedzina badań. Podgrupy, które mają wysoki wskaźnik połączeń wewnętrznych, ale niski wskaźnik połączeń zewnętrznych, to obszar badawczy niezależnego rozwoju (ang. *growth research area*), a podgrupy, w których oba wskaźniki są niskie, to obszary badawcze następnej generacji (ang. *next generation research area*), które nie mają jeszcze struktury i są badane osobno.

I wreszcie podgrupy, które mają niski wskaźnik połączeń wewnętrznych, ale wysoki wskaźnik połączeń zewnętrznych, to obszar badawczy trendów (ang. *trend research area*), który cechuje się dużą skalowalnością.

Celem wykonania tej pracy było poznanie struktury wiedzy w badaniach nad przemysłem stalowym, przy użyciu analizy sieciowej Big Data. W tym celu wykorzystano publikacje ukazujące się na łamach „Steel Research International”. Wyniki analizy pozwalają nam lepiej zrozumieć zebrane dotychczas opracowania badań. Co więcej, czytelne wydzielenie obszarów na podstawie podgrup może być cenną wskazówką dotyczącą kierunku kolejnych badań. Konieczne jest jednak przeprowadzenie bardziej wyczerpujących badań na ten temat, z uwzględnieniem zawartości innych branżowych periodyków, co wykracza poza ramy tej pracy.

### Przypisy

- 1 Jest to akronim utworzony od nazwisk twórców tego algorytmu: Clauset, Newman i Moore (przyp. tłum.).

### Literatura

- [1] LEE H., SOHN I.: *Looking back at Steel Research International and its future*. „Steel Research International”, 86(1)/2015.

Fragment pochodzi z książki:

*Big Data w przemyśle*

Lee Hyunjoung, Sohn Il

Wydawnictwo Naukowe PWN, 2019