POLSKA AKADEMIA NAUK
Instytut Badań Systemowych

# ROZMYTOŚĆ I BIPOLARNOŚĆ W INTELIGENTNYM WYSZUKIWANIU INFORMACJI

Sławomir Zadrożny

Warszawa 2013


**POLSKA AKADEMIA NAUK**
**INSTYTUT BADAŃ SYSTEMOWYCH**

**Seria: BADANIA SYSTEMOWE**
**Tom 73**

**Redaktor naukowy:**
**Prof. dr hab. inż. Jakub Gutenbaum**

**Warszawa 2013**

Rada redakcyjna serii: BADANIA SYSTEMOWE


**POLSKA AKADEMIA NAUK**
**INSTYTUT BADAŃ SYSTEMOWYCH**

**Sławomir Zadrożny**

# ROZMYTOŚĆ I BIPOLARNOŚĆ W INTELIGENTNYM WYSZUKIWANIU INFORMACJI

**Warszawa 2013**

## Autorzy:

**Dr hab. Sławomir Zadrożny**
Instytut Badań Systemowych Polskiej Akademii Nauk
ul. Newelska 6, 01-447 Warszawa
*Slawomir.Zadrozny@ibspan.waw.pl*

## Recenzenci:

**dr hab. inż. Maciej Krawczak**
**dr Marek Reformat**

**Skład:** Aneta M. Pielak

**Wydawca:**

**Instytut Badań Systemowych**
**Polskiej Akademii Nauk**
Newelska 6, 01-447 Warszawa
www.ibspan.waw.pl

**ISSN 0208-8029**
**ISBN 83-894-7551-0**

# Rozdział 6

# Wyszukiwanie informacji tekstowej – wprowadzenie

*Dziedzina wyszukiwania informacji tekstowej* (ang. *information retrieval, IR*) [216, 199, 5, 163, 76, 141] zajmuje się zagadnieniami związanymi z szeroko rozumianą reprezentacją i przetwarzaniem informacji tekstowej. Podstawowym zadaniem rozważanym w ramach tej dziedziny jest wyszukiwanie dokumentów spełniających *potrzeby informacyjne* użytkownika wyrażone w postaci *zapytania*. Dokumenty takie określa się mianem *dokumentów relewantnych* względem zapytania.

*Relewantność* (ang. *relevance*) jest jednym z najważniejszych pojęć dziedziny wyszukiwania informacji tekstowej. Ma ono charakter wysoce subiektywny: relewantne są te dokumenty, które spełniają potrzeby informacyjne *danego użytkownika* względem *danego zapytania*. Klasycznie traktowano to pojęcie jako binarne: dokument jest relewantny bądź nierelewantny. Zazwyczaj jednak przyjmuje się, że ma ono charakter *stopniowalny*: jeden dokument może być bardziej relewantny niż inny. W literaturze często stosuje się pojęcie *stopnia dopasowania* (ang. *matching degree* lub *RSV: return status value*), który wyraża stopień/intensywność tak rozumianej relewantności. Przyjmując przedział [0,1] jako zbiór wartości stopni relewantności, w naturalny sposób można zinterpretować zbiór dokumentów relewantnych jako *zbiór rozmyty*, a stopień relewantności jako *wartość funkcji przynależności tego zbioru* dla danego dokumentu. Podstawowym zadaniem systemu wyszukiwania informacji tekstowej jest wybranie z danej kolekcji dokumentów jak największej liczby dokumentów relewantnych przy jednoczesnym pominięciu dokumentów nierelewantnych. Rozwiązanie tego zadania klasycznymi metodami IR wymaga, w pierwszej kolejności, wyznaczenia pewnej uproszczonej re-

prezentacji dokumentów oraz zapytań, zazwyczaj w postaci kombinacji *słów kluczowych*, czyli słów *ważnych* z punktu widzenia znaczenia dokumentu. Proces tworzenia reprezentacji, zwany *indeksowaniem* jest zazwyczaj realizowany automatycznie przez komputer. Realizacja "ręczna" przez człowieka może dać lepsze wyniki, ale jest zazwyczaj praktycznie niewykonalna ze względu na rozmiary rozpatrywanych kolekcji dokumentów.

Przyjmowana w praktyce reprezentacja dokumentu zazwyczaj jedynie w przybliżeniu oddaje jego treść. Podobnie, przyjęta forma zapytania pozwala zwykle jedynie na niedoskonałą reprezentację potrzeb informacyjnych użytkownika. Dodatkowo, często reprezentacja taka odwołuje sie do pewnych wielkości liczbowych wyrażających istotność danego słowa kluczowego czy pojęcia. Z tych względów reprezentacja dokumentów i zapytań jest siłą rzeczy *nieprecyzyjna*, zaś cały proces przetwarzania obarczony jest *niepewnością*. Obie te formy niedoskonałości informacji są w różnym stopniu bezpośrednio uwzględniane w modelach wyszukiwania informacji tekstowej. W literaturze można wyróżnić dwa główne nurty badań w tym zakresie odnoszące się, odpowiednio, do probabilistyki [190, 108, 210, 191, 217, 17, 192, 190, 70, 109, 209, 211] i szeroko rozumianej logiki rozmytej [18, 145, 146, 144, 23, 24, 26, 143, 29, 174, 28, 185, 186]. W p. 7.2 przedstawimy szerzej jeden z modeli, którego głównym celem jest jednoczesne uwzględnienie nieprecyzyjności i niepewności w ramach jednego formalizmu bazującego na szeroko rozumianej logice rozmytej.

## 6.1 Modele IR

W literaturze zaproponowano wiele podejść do wyszukiwania informacji tekstowej, nazywanych *modelami wyszukiwania informacji tekstowej*. Różnią się one między sobą rozwiązaniami odnośnie do:

1. reprezentacji dokumentów,
2. reprezentacji zapytań,
3. sposobu określania dopasowania dokumentów i zapytań.

Zwykle w literaturze wyróżnia się trzy podstawowe modele wyszukiwania informacji, zwane również modelami klasycznymi [5]:

- boolowski (logiczny)
- wektorowy

- probabilistyczny

W literaturze zaproponowano wiele ich rozwinięć. Opiszemy teraz pokrótce wybrane modele, istotne z punktu widzenia możliwości zastosowania aparatu logiki rozmytej. Rozpoczniemy od krótkiego przedstawienia bazowych wersji modeli klasycznych. Szerszy przegląd modeli można znaleźć w pracach [5, 216, 199, 142, 9, 160, 192, 108, 148, 149, 62].

W dalszej części przyjmiemy następujące oznaczenia:

$K = \{k_j\}_{j \in J}$ zbiór słów kluczowych używanych do indeksowania dokumentów,

$D = \{d_l\}_{l \in L}$ zbiór dokumentów tekstowych w rozważanej kolekcji.

### 6.1.1 Model boolowski i jego rozszerzenia

Poszczególne elementy *modelu boolowskiego* można następująco wyrazić w terminach logiki klasycznej, a dokładniej rachunku zdań (por. p. 2.2).

**Reprezentacja dokumentów.** Dokument reprezentowany jest jako zbiór przypisanych mu słów kluczowych. Nie rozróżnia się stopni ważności słów kluczowych. W szczególności częstość występowania słów w dokumencie nie jest brana pod uwagę.

W języku rachunku zdań reprezentację dokumentów można opisać w następujący sposób. Z każdym słowem kluczowym $k_i \in K$ wiążemy zmienną zdaniową $s_i \in S$ (por. s. 27). Dokument $d$ traktujemy jako *wartościowanie* $\omega_d$ takie że:

$$\omega_d(s_i) = \begin{cases} 1 & \text{jeśli słowo kluczowe } k_i \text{ jest przypisane dokumentowi } d \\ 0 & \text{w przeciwnym przypadku} \end{cases} \tag{6.1}$$

Przypisanie zmiennej zdaniowej $s_i$ w powyższej reprezentacji wartości 0 (fałsz) stanowi przyjęcie tzw. *założenia o zamkniętości świata* (CWA)[1]: przypuszczamy, że to o czym *nie wiemy* że jest prawdą, jest fałszem. Możliwe jest też inne podejście, przyjęcie tzw. *założenia o otwartości świata* (OWA)[2], przy którym takiego przypuszczenia nie czynimy. Przyjęcie OWA skutkuje zastosowaniem do reprezentacji dokumentów "nie-

[1] ang. *Closed World Assumption (CWA)*
[2] ang. *Open World Assumption (OWA)*

kompletnego wartościowania", które zmiennym zdaniowym odpowiadającym słowom kluczowym niewystępującym w dokumencie nie przypisuje żadnej wartości logicznej[3]. To rozróżnienie ma znaczenie w niektórych podejściach omawianych w dalszej części książki.

Warto zwrócić uwagę, że wyżej omówiona reprezentacja odpowiada przypisaniu dokumentowi formuły $\phi$ będącej koniunkcją:

$$\phi = s_1 \wedge s_2 \wedge \ldots \wedge s_n \tag{6.2}$$

lub

$$\phi = s_1 \wedge s_2 \wedge \ldots \wedge s_n \wedge \neg s_{n+1} \wedge \neg s_{n+2} \wedge \ldots \wedge s_m \tag{6.3}$$

gdzie, jak wcześniej, zmienne zdaniowe $s_i$ odpowiadają poszczególnym słowom kluczowym $k_i$, użytym (zmienne $s_1 - s_n$) lub nieużytym (zmienne $s_{n+1} - s_m$), do reprezentacji dokumentu. Wartościowanie (6.1) jest jedynym modelem formuły (6.3) i jednym z modeli formuły (6.2). Jeśli jednak przyjmiemy CWA, to obydwie formuły stają się rownoważne i wartościowanie (6.1) jest ich jedynym wspólnym modelem. Jednocześnie można rozważyć ustalenie jako reprezentacji dokumentu bardziej złożonej formuły, skonstruowanej z użyciem różnych spójników logicznych. Wtedy reprezentacja opisana przypisaniem (6.1) łatwo poddaje się uogólnieniu: dokument, któremu przypisano formułę $\phi$ będzie reprezentowany przez zbiór jej modeli $\Omega^{\phi}$ (por. (2.55)).

Taka złożona formuła reprezentująca dokument może na przykład przyjąć *dysjunktywną postać normalną* lub *konjunktywną postać normalną*:

- $(s_{11} \wedge s_{12} \wedge \ldots s_{1l_1}) \vee (s_{21} \wedge s_{22} \wedge \ldots s_{2l_2}) \vee \ldots \vee (s_{n1} \wedge s_{n2} \wedge \ldots s_{nl_n})$

- $(s_{11} \vee s_{12} \vee \ldots s_{1l_1}) \wedge (s_{21} \vee s_{22} \vee \ldots s_{2l_2}) \wedge \ldots \wedge (s_{n1} \vee s_{n2} \vee \ldots s_{nl_n})$

Pierwsza z powyższych formuł może być użyteczna przy reprezentowaniu dokumentu składającego się z kilku wyróżnionych części, przy czym każda z tych części reprezentowana jest przez koniunkcję słów kluczowych. Druga z powyższych formuł może być dogodnym rozszerzeniem bazowej koniunkcji (6.2), powstałym przez uwzględnienie synonimów poszczególnych słów kluczowych: zakładamy, że zmienne zdaniowe $\{s_{ij}\}_{j=1,l_i}$ reprezentują słowa kluczowe, które są synonimami.

[3]Formalnie rzecz biorąc, nie jest to wtedy wartościowanie w sensie (2.51) i stąd używamy tu cudzysłowu.

**Reprezentacja zapytań.** Zapytanie wyrażone jest w postaci formuły logicznej $\phi \in \Phi$ (będziemy też używać oznaczenia $q$). Formuła ta konstruowana jest z użyciem zmiennych zdaniowych, które odpowiadają słowom kluczowym reprezentującym *potrzeby informacyjne* użytkownika. Na przykład potrzeby informacyjne użytkownika zainteresowanego miastami we Francji i Anglii można wyrazić zapytaniem o postaci następującej formuły $\phi$:

$$\phi = s_1 \wedge (s_2 \vee s_3)$$

lub, bardziej bezpośrednio, w formie zapytania do wyszukiwarki internetowej:

```
miasto AND (Francja OR Anglia)
```

**Ocena dopasowania dokumentu i zapytania.** Ocena dopasowania (relewantności) dokumentu $d$ względem zapytania $q$ utożsamiana jest z prawdziwością $\omega_d(q)$ formuły logicznej $q$, reprezentującej zapytanie, przy wartościowaniu $\omega_d$ określonym przez dokument (por. (6.1)). W ramach klasycznego rachunku zdań ocena ta jest binarna: formuła logiczna może być jedynie prawdziwa lub fałszywa. W klasycznym modelu boolowskim dokument może być jedynie uznany za relewantny bądź nierelewantny.

Założenie o binarności relewantności jest nieadekwatne: użytkownik zazwyczaj będzie skłonny wyróżnić wiele poziomów relewantności i stwierdzić na przykład, że pewien dokument $d_1$ jest *bardziej relewantny* niż dokument $d_2$, który z kolei jest *bardziej relewantny* niż $d_3$, przy czym $d_3$ nadal zostanie przez niego uznany za *w pewnym stopniu relewantny*. Brak możliwości uporządkowania w ten sposób dokumentów jest najważniejszą wadą modelu boolowskiego. W odpowiedzi na zapytanie użytkownik otrzymuje *zbiór* dokumentów uznanych za relewantne, zamiast postulowanego wcześniej *zbioru uporządkowanego*, określającego ranking dokumentów.

**Przykład 6.1.** *Rozważmy zbiór dokumentów $D = \{d_l\}_{l=1,2,3}$ indeksowanych z użyciem słów kluczowych z następującego zbioru $K = \{k_j\}_{j=1,\ldots,6}$. Niech poszczególne dokumenty będą reprezentowane przez następujące zbiory słów kluczowych:*

$$d_1 = \{k_1, k_2, k_3\}, \quad d_2 = \{k_3, k_5\}, \quad d_3 = \{k_5, k_6\}$$

*Dokumentom tym odpowiadają więc następujące wartościowania:*

$$\omega_{d_1}(s_1) = \omega_{d_1}(s_2) = \omega_{d_1}(s_3) = 1, \quad \omega_{d_1}(s_4) = \omega_{d_1}(s_5) = \omega_{d_1}(s_6) = 0$$

$$\omega_{d_2}(s_3) = \omega_{d_2}(s_5) = 1, \quad \omega_{d_2}(s_1) = \omega_{d_2}(s_2) = \omega_{d_2}(s_4) = \omega_{d_2}(s_6) = 0$$
$$\omega_{d_3}(s_5) = \omega_{d_3}(s_6) = 1, \quad \omega_{d_3}(s_1) = \omega_{d_3}(s_2) = \omega_{d_3}(s_3) = \omega_{d_3}(s_4) = 0$$

*Rozważmy teraz dwa zapytania postaci:*

$$q_1 = s_1 \vee s_6, \quad q_2 = s_5 \wedge s_6$$

*Wtedy:*

$$\omega_{d_1}(q_1) = \max(1,1) = 1, \quad \omega_{d_1}(q_2) = \min(0,0) = 0$$
$$\omega_{d_2}(q_1) = \max(0,0) = 0, \quad \omega_{d_2}(q_2) = \min(1,0) = 0$$
$$\omega_{d_3}(q_1) = \max(0,1) = 1, \quad \omega_{d_3}(q_2) = \min(1,1) = 1$$

Tak więc odpowiedziami na dwa rozważane zapytania są następujące zbiory dokumentów:

$$q_1 \longmapsto \{d_1, d_3\}$$
$$q_2 \longmapsto \{d_3\}$$

Mimo wspomnianej wady język zapytań modelu boolowskiego jest bardzo bogaty. Pozwala on na wyrażenie złożonych potrzeb informacyjnych użytkownika. Jednocześnie posługiwanie się tym językiem w sposób efektywny może być trudne dla osób pozbawionych przygotowania matematyczno-informatycznego.

### Modele logiczne

Model boolowski jest podstawowym wariantem całej rodziny *modeli logicznych*. Podstawą modeli logicznych jest reprezentowanie dokumentów i zapytań w postaci formuł pewnej logiki lub ich pewnych równoważnych form (por. (6.1)). W modelu boolowskim stosuje się klasyczny rachunek zdań, co nie pozwala na uwzględnienie stopniowalności ważności słów kluczowych dla reprezentacji dokumentu czy zapytania ani stopniowalności pojęcia relewantności. Nie ma również możliwości uwzględnienia niepewności i nieprecyzyjności związanych z reprezentacją dokumentów i zapytań oraz określaniem ich dopasowania.

W literaturze można znaleźć wiele podejść zmierzających do uelastycznienia modelu boolowskiego pod tym względem. Najbardziej znanym podejściem tradycyjnym jest model $p$-norm.

Model ten został zaproponowany w latach 80. ubiegłego wieku przez Saltona, Foxa i Wu [198]. Stanowi on rozwinięcie modelu boolowskiego

poprzez wprowadzenie *stopni ważności* (*wag*) słów kluczowych i uwzględnienie możliwości częściowego dopasowania dokumentów względem zapytań. W modelu $p$-norm przyjmuje się stopnie ważności wyłącznie przy określaniu reprezentacji dokumentów. Zapytania mają taką samą postać jak w modelu boolowskim.

Rozpatrzmy zapytanie w postaci koniunkcji dwóch słów kluczowych $k_1$ i $k_2$ (dla uproszczenia utożsamiamy tu symbol $k_i$, oznaczający słowa kluczowe, z symbolem oznaczającym przypisaną mu zmienną zdaniową – por. p. 6.1.1):

$$q_\wedge = k_1 \wedge k_2 \tag{6.4}$$

Do określenia dopasowania dokumentów do tego zapytania istotne są jedynie wagi przypisane w tych dokumentach słowom kluczowym $k_1$ i $k_2$. Zakładając, że wagi te są liczbami z przedziału [0,1] dogodnie jest przedstawić dokumenty jako wektory w przestrzeni $R^2$. Dokument tym lepiej spełnia zapytanie (6.4) im wyższe wagi mają w nim słowa kluczowe $k_1$ i $k_2$. W modelu $p$-norm przyjmuje się więc, że stopień spełnienia jest funkcją odległości wektora reprezentującego dokument od wektora [1,1]. Odległość wektorów standardowo definiuje się jako normę wektora będącego ich różnicą. Zależnie od przyjętej postaci normy ($p = 1, 2, \infty$) otrzymuje się różne wzory na stopień dopasowania. Na przykład dla normy euklidesowej wzór przyjmuje następującą postać, wyrażającą dopełnienie do jedności znormalizowanej odległości wektorów (czyli ich bliskości):

$$MD(q_\wedge, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}} \tag{6.5}$$

gdzie $MD$ oznacza stopień dopasowania, zaś dokument $d$ reprezentowany jest przez wektor $[x, y]$ stopni ważności słów kluczowych $k_1$ i $k_2$.

Rozpatrzmy teraz zapytanie w postaci alternatywy dwóch słów kluczowych $k_1$ i $k_2$ czyli:

$$q_\vee = k_1 \vee k_2 \tag{6.6}$$

W naturalny sposób można przyjąć, że dokument tym lepiej spełnia zapytanie im większa jest odległość wektora go reprezentującego $[x, y]$ od wektora $[0, 0]$ odpowiadającego dokumentowi, w reprezentacji którego obydwa słowa kluczowe $k_1$ i $k_2$ nie odgrywają żadnej roli. Wzór na stopień dopasowania (odległość) przyjmuje w tym przypadku następującą postać (znów zakładając użycie normy euklidesowej):

$$MD(q_\vee, d) = \sqrt{\frac{x^2 + y^2}{2}} \tag{6.7}$$

Zastosowanie modelu $p$-norm pozwala przezwyciężyć pewne ograniczenia klasycznego modelu logicznego: brak możliwości różnicowania ważności słów kluczowych przy reprezentacji dokumentów oraz brak możliwości uszeregowania dokumentów w odpowiedzi na zapytanie (poza prostym odróżnieniem relewantnych od nierelewantnych). Obie wady wynikają z binarnego charakteru logiki przyjętej w klasycznym modelu logicznym.

Losada i Barreiro [156] proponują podejście w ramach klasycznego modelu boolowskiego, w którym dopasowanie pomiędzy dokumentem $d$ a zapytaniem $q$ określa się na podstawie odpowiednio zdefiniowanej średniej odległości pomiędzy zbiorami modeli (2.55) formuł reprezentujących dokument i zapytanie. Pozwala to odejść od binarnego pojmowania relewantności dokumentu względem zapytania. Jednak nadal ważność poszczególnych słów dla reprezentacji dokumentów i zapytań traktowana jest binarnie. Ci sami autorzy zaproponowali [157] rozszerzenie swojego podejścia, pozwalające złagodzić nieco również to ograniczenie. Zasadniczo, podejście to polega na tym, że przy określaniu odległości pomiędzy modelami nie bierze się pod uwagę wyłącznie liczby zmiennych zdaniowych, którym porównywane modele (wartościowania) przypisują różne wartości prawdy, jak to ma miejsce w przypadku wcześniejszego podejścia [156]. W rozszerzonym podejściu bierze się również pod uwagę podobieństwo słów kluczowych oraz częstość ich występowania w kolekcji dokumentów. Dokładniej, jeśli w dokumencie nie występuje słowo kluczowe, które występuje w zapytaniu, to to niedopasowanie wpłynie słabiej na różnicę pomiędzy nimi, jeśli w dokumencie występuje inne, ale podobne słowo kluczowe. Zakłada się, że na zbiorze słów kluczowych $K$ określona jest rozmyta relacja podobieństwa (rodzaj rozmytego tezaurusa), która określa podobieństwa pomiędzy wszystkimi parami słów kluczowych. Częstość występowania słowa kluczowego w kolekcji mierzona jest wskaźnikiem IDF (por. p. 6.1.2) i im ta częstość wyższa tym współwystępowanie tego słowa w dokumencie i zapytaniu mniej wpływa na ich podobieństwo (podobieństwo reprezentujących je modeli). Intuicja uzasadniająca takie postępowanie jest dość oczywista - te słowa, które w ogólności występują często nie mają decydującego znaczenia przy określaniu podobieństwa zapytania i dokumentu, natomiast te które występują rzadko i wystąpią jednocześnie w zapytaniu i w dokumencie faktycznie wskazują silniej na ich podobieństwo/dopasowanie. Podejście Losady i Barreiro nie jest samo w sobie szczególnie interesujące dla naszych rozważań w niniejszej książce. Obejmuje ono jednak elementy stosowane w podejściu Liau i Yao, które opisujemy w dalszej części rodziału, i które wprost odnosi się do szeroko rozumianej logiki rozmytej.

W literaturze znaleźć można pewną liczbę prac dotyczących zastosowania teorii możliwości (por. 2.1.2) do wyszukiwania informacji tekstowej. Szczególnie interesująca jest praca Liau i Yao [150], gdzie również zakłada się, że dokumenty i zapytania reprezentowane są przez zbiory modeli (2.55) reprezentujących je formuł, podobnie jak w pracach Losady i Barreiro. Podobnie, definiują oni i stosują znormalizowaną odległość pomiędzy modelami formuł reprezentujących zapytania i dokumenty. Przyjmuje się, że każdy dokument $d$ generuje rozkład możliwości w przestrzeni wartościowań $\Omega$, taki że stopień możliwości danego wartościowania $\omega$ jest równy stopniowi jego podobieństwa do wartościowania $\omega_d$ (6.1). Stopień dopasowania dokumentu i zapytania określony jest przez parę wartości miar możliwości i konieczności (2.29)-(2.33) obliczonych dla zbioru modeli formuły reprezentującej zapytanie. Liau i Yao wyróżniają dwa sposoby reprezentacji dokumentów, odpowiadające CWA i OWA (por. s. 131). Wspólnie dla obu sposobów zakłada się, że na zbiorze dokumentów $D$ określona jest relacja podobieństwa $sim$. Równoważnie, zakłada się istnienie relacji podobieństwa $sim$ na zbiorze wszystkich wartościowań $\Omega$ określonych na alfabecie $S = \{s_i\}$. Taka relacja podobieństwa może być dana lub określa się ją z użyciem odległości na wartościowaniach, zdefiniowanej tak samo jak u Losady i Barreiro.

Najpierw rozpatrzmy podejście oparte na CWA. Na podstawie relacji podobieństwa $sim$ dla każdego dokumentu $d$ określa się rozkład możliwości $\pi_d$ na zbiorze wszystkich wartościowań $\Omega$ o następującej postaci:

$$\pi_d(\omega) = sim(\omega^d, \omega)$$

gdzie $\omega, \omega^d \in \Omega$ i $\omega^d$ tak jak poprzednio jest *modelem* $d$ – przy założeniu CWA taki model istnieje tylko jeden.

Stopień dopasowania $MD(d, q)$ pomiędzy dokumentem $d$ i zapytaniem $q$ określa się jako parę:

$$MD(d, q) = (\Pi_d(\Omega^q), N_d(\Omega^q)) \tag{6.8}$$

wartości miar możliwości i konieczności zbioru modeli $q$, obliczonych z użyciem rozkładu mozliwości $\pi_d$:

$$\Pi_d(\Omega^q) = \max_{\omega_i \in \Omega^q} \pi_d(\omega_i)$$

$$N_d(\Omega^q) = \min_{\omega_i \notin \Omega^q} (1 - \pi_d(\omega_i))$$

Relacja porządku $\prec_q$ określona na zbiorze dokumentów $D$ względem spełniania przez nie zapytania $q$ zdefiniowana jest jako porządek leksykograficzny na parach (6.8):

$$d_1 \prec_q d_2 \Leftrightarrow \Pi_{d_1}(\Omega^q) < \Pi_{d_2}(\Omega^q) \vee \Pi_{d_1}(\Omega^q) = \Pi_{d_2}(\Omega^q) \wedge N_{d_1}(\Omega^q) < N_{d_2}(\Omega^q) \tag{6.9}$$

W przypadku przyjęcia OWA dokument $d$ ma w ogólnosci zbiór modeli $\Omega^d$. Należy to uwzglednic w formule (6.8). Liau i Yao proponują dwa warianty $MD_1$ (optymistyczny) i $MD_2$ (pesymistyczny):

$$MD_1(d,q) = \max_{>_{lex}}\{MD(\omega,q) : \omega \in \Omega^d\} \tag{6.10}$$

$$MD_2(d,q) = \min_{>_{lex}}\{MD(\omega,q) : \omega \in \Omega^d\} \tag{6.11}$$

gdzie $>_{lex}$ oznacza porządek leksykograficzny.

Inny interesujący nurt prac [50, 51], związany jest z zastosowaniem tak zwanych sieci posybilistycznych.

### 6.1.2 Model wektorowy

W modelu wektorowym dokumenty i zapytania reprezentowane są jako wektory w wielowymiarowej przestrzeni wektorowej. Każdemu słowu kluczowemu $k_j \in K$ odpowiada wymiar przestrzeni. Współrzędne wektorów reprezentujących dokumenty i zapytania to stopnie ważności przypisane poszczególnym słowom kluczowym.

**Reprezentacja dokumentów.** Reprezentację dokumentu $d_i$ stanowi wektor skonstruowany na podstawie przypisanych mu słów kluczowych. Dla każdego słowa kluczowego $k_j$ określony jest stopień ważności (waga) $w_{ij}$ w dokumencie $d_i$, który stanowi $j$-tą współrzędną wektora. Często wagi słów kluczowych normalizuje się tak, że współrzędne wektorów przyjmują wartości z przedziału [0,1] lub norma wektora jest równa 1. Zakłada się, że przy takiej reprezentacji wektory reprezentujące podobne tematycznie dokumenty będą sobie bliskie w sensie pewnej miary odległości. Wektor dla dokumentu $d_i$ można zapisać następująco:

$$d_i = [w_{i1}, w_{i2}, \ldots, w_{i|K|}]$$

gdzie $K$, jak poprzednio, jest zbiorem wszystkich słów kluczowych, zaś $|K|$ oznacza jego liczność.

**Reprezentacja zapytań.** Zapytania użytkownika $q$ są przedstawiane analogicznie do dokumentów jako wektory wag $w_j$ poszczególnych słów kluczowych w nich występujących. Wektor dla zapytania definiujemy więc następująco:

$$q = \left[v_1, v_2, \ldots, v_{|K|},\right]$$

**Ocena relewantności.** Stopień dopasowania dokumentu i zapytania określa się jako podobieństwo reprezentujących je wektorów. Najpopularniejszą miarą $Sim(d, q)$ tego podobieństwa jest cosinus kąta pomiędzy tymi wektorami.

$$Sim(d_i, q) = \frac{\sum_{j=1}^{|K|} w_{ij} * v_j}{\sqrt{\sum_{j=1}^{|K|} w_{ij}^2} * \sqrt{\sum_{j=1}^{|K|} v_j^2}} \tag{6.12}$$

Ocena relewantności jest stopniowalna. Zero oznacza całkowity brak dopasowania, 1 całkowite dopasowanie, a wartości pośrednie częściowe dopasowanie. W odpowiedzi na zapytanie dokumenty mogą być uporządkowane nierosnąco względem ich podobieństwa do zapytania.

Skuteczność działania systemu opartego na modelu wektorowym w dużej mierze zależy od właściwego doboru wag poszczególnych słów kluczowych w reprezentacji dokumentów i zapytań. Opracowano wiele metod (*schematów ważenia*) określania tych wag. Szeroki przegląd stosowanych schematów zawiera praca Saltona i Buckleya [197]. Przyjmuje się, że każdy schemat składa się z trzech składowych, uwzględniających dla poszczególnych słów kluczowych:

(i) *częstość występowania słowa w dokumencie lub zapytaniu*[4]; im większa częstość występowania słowa w dokumencie, tym większa wartość tej składowej wagi; wartość tej składowej nie musi być wprost równa częstości słowa kluczowego w dokumencie. Może być ona binarna: równa 1, jeśli słowo występuje i 0 w przeciwnym wypadku; wartość ta może być również normalizowana; składowa ta może mieć w każdym dokumencie inną wartość dla tego samego słowa kluczowego.

(ii) *częstość występowania słowa w zbiorze dokumentów*[5]; jest to wartość odwrotnie proporcjonalna do liczby dokumentów w zbiorze $D$, w których występuje dane słowo kluczowe; największą wartość

[4] ang. *term frequency, tf*
[5] ang. *collection/document frequency*

tej składowej wagi uzyskują te słowa kluczowe, które występują w niewielu dokumentach – traktowane są one jako bardziej znaczące dla reprezentacji zawartości poszczególnych dokumentów niż słowa kluczowe występujace w większości dokumentów w kolekcji; wartość tej składowej jest więc identyczna dla danego słowa kluczowego we wszystkich dokumentach zbioru $D$.

(iii) *normalizację wektora*[6]; wektory reprezentujące dokumenty poddaje sie normalizacji w celu równorzędnego traktowania dokumentów różnej długości – bez jej zastosowania dokumenty dłuższe będą miały zazwyczaj dużo większe wartości poszczególnych współrzędnych wektorów je reprezentujących.

Przykładowe, popularniejsze formy poszczególnych składowych wag to:

(i) *tf*, liczba wystąpień danego słowa w dokumencie lub zapytaniu (ang. *term frequency*),

(ii) *IDF*, $\log \frac{N}{n}$, gdzie $N$ to liczba wszystkich dokumentów w zbiorze, zaś $n$ to liczba dokumentów, w których występuje dane słowo kluczowe (ang. *inverse document frequency*),

(iii) normalizacja z użyciem normy euklidesowej wektora: $w_{ij} = \frac{w_{ij}}{\sqrt{\sum_j w_{ij}^2}}$.

Najczęściej stosowanym w literaturze schematem ważenia (kombinacją tych trzech składowych) jest schemat oznaczany jako *tf*IDF*[7], który wyraża się następującym łącznym wzorem:

$$w_{ij} = tf_{ij} * \log\left(\frac{N}{n_j}\right) \tag{6.13}$$

gdzie jak poprzednio $w_{ij}$ jest wagą słowa kluczowego $k_j$ w dokumencie $d_i$; $tf_{ij}$ jest częstością występowania słowa kluczowego $k_j$ w dokumencie $d_i$; $N$ jest liczbą wszystkich dokumentów w zbiorze $D$, a $n_j$ jest liczbą dokumentów, w których wystąpiło słowo $k_j$. Zazwyczaj dodatkowo stosowana jest normalizacja tak określonego wektora wag z użyciem normy euklidesowej.

Ta najpopularniejsza metoda obliczania wag słów kluczowych przypisuje słowu $k_j$ w dokumencie $d_i$ wagę proporcjonalną do liczby wystąpień

[6] ang. *normalization*

[7] ang. *term frequency–inverse document frequency*

tego słowa w dokumencie $d_i$ i odwrotnie proporcjonalną do liczby dokumentów w zbiorze $D$, w których słowo to pojawiło się chociaż raz.

Mocne strony modelu wektorowego to uwzględnienie zróżnicowanej ważności słów kluczowych oraz w naturalny sposób uzyskana stopniowalność oceny relewantności dokumentów. Słabe strony rozwiązania opartego na modelu wektorowym to pewna arbitralność schematów ważenia słów i pewna „sztywność" języka zapytań.

Model wektorowy stosowany jest w różnych zadaniach związanych z przetwarzaniem dokumentów tekstowych. Ważną rolę odgrywa w nich często pojęcie macierzy *słowo kluczowe-dokument* (ang. *term-document matrix*). Jest to macierz, której kolumny stanowią wektory reprezentujące poszczególne dokumenty w rozpatrywanej kolekcji. Na macierzy tej wykonywane są różne operacje, które pozwalają na przykład uzyskać wgląd w zależności zachodzące pomiędzy wystąpieniami poszczególnych słów kluczowych. Warto zwrócić uwagę, że macierz ta ma zazwyczaj bardzo duże rozmiary: zarówno liczba dokumentów, jak i liczba słów kluczowych są bardzo duże w nietrywialnych zastosowaniach. Jednocześnie macierz taka jest zazwyczaj bardzo rzadka: w pojedynczym dokumencie występuje zwykle jedynie niewielki podzbiór zbioru słów kluczowych użytych do indeksowania całej kolekcji.

### 6.1.3 Model probabilistyczny

Istotą modelu probabilistycznego jest interpretacja oceny relewantności dokumentu względem zapytania w terminach teorii prawdopodobieństwa. Dokładniej, stopień dopasowania stanowi oszacowanie prawdopodobieństwa tej relewantności. Najpełniejszy efekt uzyskuje się przy założeniu iteracyjności procesu wyszukiwania: użytkownik formułuje zapytanie, system udziela wstępnej odpowiedzi, która podlega ocenie użytkownika, na tej podstawie system generuje kolejną wersję odpowiedzi i cały proces ulega powtórzeniu. Ocena użytkownika przekazywana systemowi w kolejnych iteracjach stanowi podstawę do lepszego zidentyfikowania preferencji użytkownika i co za tym idzie lepszego oszacowania prawdopodobieństwa relewantności poszczególnych dokumentów względem zapytania. Charakterystykę tego modelu można przedstawić w następujący sposób.

**Reprezentacja dokumentów.** Dokumenty reprezentowane są jako zdarzenia elementarne $d = (w_1, \ldots, w_{|K|})$ w wielowymiarowej przestrzeni probabilistycznej. Poszczególne wymiary odpowiadają słowom kluczo-

wym używanym do indeksowania dokumentów. Każde zdarzenie elementarne odpowiada pewnej kombinacji występowania i niewystępowania poszczególnych słów kluczowych, przy czym $w_i = 1$ i $w_i = 0$ oznaczają, odpowiednio, że słowo $k_i$ występuje i nie występuje w dokumencie. Efektywnie, dokument reprezentowany jest jako zbiór przypisanych mu słów kluczowych. Pod tym względem reprezentacja dokumentów w modelu probabilistycznym jest analogiczna do przyjętej w modelu boolowskim.

**Reprezentacja zapytań.** Zapytania reprezentowane są również jako zbiory słów kluczowych i jednocześnie jako zdarzenia w przestrzeni probabilistycznej.

**Ocena relewantności.** Stopień dopasowania dokumentu i zapytania określa się jako prawdopodobieństwo relewantności dokumentu względem zapytania. Relewantność jest traktowana jako kategoria binarna, a więc jednocześnie uzyskuje się oszacowanie tego, że dokument nie jest relewantny. W odniesieniu do przyjętej postaci przestrzeni probablistycznej należy zauważyć, że dla ustalonego zapytania $q$ rozważamy faktycznie zdarzenia elementarne postaci $(d, Rel)$, gdzie $Rel$ jest zmienną binarną: $Rel = 1$ oznacza, ze dokument $d$ jest relewantny względem zapytania $q$, zaś $Rel = 0$ oznacza, że nie jest on relewantny. Kryterium wyboru dokumentów potencjalnie interesujących dla użytkownika można więc wyrazić następująco:

$$P(Rel = 1 \mid d) > P(Rel = 0 \mid d) \tag{6.14}$$

Przyjmując pewne założenia o zależności relewantności od postaci dokumentu (względem danego zapytania) i stosując do wzoru (6.14) standardowe przekształcenia uzyskuje się efektywną regułę decyzyjną. Poszczególne kroki można w skrócie przedstawić następująco. Stosując regułę Bayesa do $P(Rel = x \mid d)$ uzyskuje się:

$$P(Rel = x \mid d) = \frac{P(d \mid Rel = x)P(Rel = x)}{P(d)}$$

Po podstawieniu do wzoru (6.14) otrzymuje się:

$$\frac{P(d \mid Rel = 1)P(Rel = 1)}{P(d)} > \frac{P(d \mid Rel = 0)P(Rel = 0)}{P(d)}$$

czyli

$$P(d \mid Rel = 1)P(Rel = 1) > P(d \mid Rel = 0)P(Rel = 0)$$

a po zlogarytmowaniu obu stron:

$$\log(P(d \mid Rel = 1)P(Rel = 1)) > \log(P(d \mid Rel = 0)P(Rel = 0))$$

i po przeniesieniu wyrazów:

$$\log(P(d \mid Rel = 1)) - \log(P(d \mid Rel = 0)) > \log(P(Rel = 0)) - \log(P(Rel = 1)) \quad (6.15)$$

gdzie prawa strona ma stałą wartość, niezależną od postaci dokumentu. Reguła (6.14) po podanych wyżej przekształceniach prowadzi więc do sformułowania następującego kryterium: im większe prawdopodobieństwo dopasowania dokumentu $d$ do zapytania[8] $q$ tym wyższa jest wartość funkcji $g(d)$ [71] o postaci:

$$g(d) = \log P(d \mid Rel = 1) - \log P(d \mid Rel = 0) \quad (6.16)$$

Warto zauważyć, że tak sformułowane kryterium uwzględnia stopniowalność dopasowania dokumentu względem zapytania i pozwala zastąpić decyzję o charakterze binarnym wyrażoną wzorem (6.14).

W dalszych przekształceniach zakłada się upraszczająco, że wystąpienia poszczególnych słów kluczowych są od siebie niezależne, zarówno w dokumentach relewantnych, jak i nierelewantnych. Przy tym założeniu prawdopodobieństwo $P(d \mid Rel = 1)$ można wyrazić jako:

$$P(d \mid Rel = 1) = P(k_1 \mid Rel = 1) \times \ldots \times P(k_k \mid Rel = 1) \quad (6.17)$$

Analogicznie postępuje się dla $P(d \mid Rel = 0)$.

Oznaczmy prawdopodobieństwa, że słowo kluczowe $k_j$, odpowiednio, wystąpi w dokumencie relewantnym i nierelewantnym następująco:

$$P_j = P(w_j = 1 \mid Rel = 1) \quad (6.18)$$

$$Q_j = P(w_j = 1 \mid Rel = 0) \quad (6.19)$$

Wtedy wzory (6.17) i (6.16) można zapisać w następującej postaci:

$$P(d \mid Rel = 1) = \Pi_{j=1}^{|K|} P_j^{w_j}(1 - P_j)^{1-w_j} \quad (6.20)$$

$$P(d \mid Rel = 0) = \Pi_{j=1}^{|K|} Q_j^{w_j}(1 - Q_j)^{1-w_j} \quad (6.21)$$

[8]Postać zapytania $q$ ma wpływ na prawdopodobieństwo warunkowe występujące w omawianych wzorach. Zostanie to explicite wyrażone w dalszej dyskusji oceny stopnia dopasowania w modelu probabilistycznym.

$$g(d) = \sum_{j=1}^{|K|} w_j \log \frac{P_j(1-Q_j)}{(1-P_j)Q_j} + \sum_{j=1}^{|K|} \log \frac{1-P_j}{1-Q_j} \tag{6.22}$$

Wartość drugiego składnika we wzorze (6.22) jest identyczna dla wszystkich dokumentów, gdyż nie odnosi się on do postaci dokumentu analizowanego pod względem dopasowania do zapytania. Można więc ten składnik pominąć w dalszej analizie, ponieważ nie wpływa on na uporządkowanie dokumentów względem wartości $g(d)$, a właśnie to uporządkowanie a nie wartości absolutne funkcji $g(d)$ jest istotne przy generowaniu odpowiedzi na zapytanie.

Pierwszy składnik (6.22) można wyrazić w dogodnej do dalszej analizy dwuskładnikowej postaci:

$$g(d) = \sum_{j} w_j \log \frac{P_j}{1-P_j} + \sum_{j} w_j \log \frac{1-Q_j}{Q_j} \tag{6.23}$$

Dodatkowo, przyjmuje się, że sumowanie we wzorze (6.23) przebiega już tylko po słowach kluczowych występujących w zapytaniu.

Aby móc efektywnie stosować wzór (6.23) jako podstawę do generowania odpowiedzi na zapytanie, należy określić sposób szacowania wartości $P_j$ i $Q_j$. Przyjmijmy, że wyszukiwanie ma charakter iteracyjny i w kolejnych krokach ma miejsce interakcja z użytkownikiem polegająca na tym, że użytkownik każdorazowo wskazuje dokumenty relewantne wśród tych przedstawianych mu przez system w kolejnych iteracjach[9]. Wtedy, szacowanie wspomnianych prawdopodobieństw może odnosić się do częstości występowania poszczególnych słów kluczowych w dokumentach wskazanych jako relewantne (szacowanie $P_j$) i pozostałe, czyli nierelewantne (szacowanie $Q_j$). Pozostaje problem pierwszej iteracji wyszukiwania, kiedy system nie posiada jeszcze informacji zwrotnej od użytkownika. Przyjmuje się wtedy, że wystąpienie wszystkich słów kluczowych w dokumentach relewantnych jest równie prawdopodobne. Przyjmuje się również, że prawdopodobieństwo $Q_j$ można oszacować jako $n_j/N$, gdzie $n_j$ to liczba dokumentów, w których słowo kluczowe $k_j$ występuje, natomiast $N$ to liczba wszystkich dokumentów w zbiorze $D$. To szacowanie jest uzasadnione założeniem, że liczba dokumentów nierelewantnych jest zazwyczaj o rząd wielkości większa od liczby dokumentów relewantnych.

[9]Można założyć, że system przedstawia użytkownikowi $n$ najlepszych dokumentów w sensie uporządkowania według (6.23).

Wtedy wzór (6.23) można przekształcić do następującej postaci:

$$C \sum_j w_j + \sum_j w_j \log \frac{N - n_j}{n_j} \tag{6.24}$$

gdzie $C$ jest stałą.

Uzyskany wzór na ocenę dopasowania dokumentu względem zapytania można zinterpretować jako kombinację prostego dopasowania określonego przez wielkość proporcjonalną do liczby słów kluczowych wspólnie występujących w zapytaniu i dokumencie (pierwszy składnik wzoru (6.24)) oraz odwrotności częstości występowania słowa w zbiorze dokumentów ($IDF$) – dla dużych $N$ drugi składnik wzoru (6.24) jest bliski postaci wzoru na $IDF$ (6.13).

Zaletą modelu jest ranking dokumentów na podstawie oceny relewantności w stosunku do zapytania oraz jego podstawy teoretyczne - zastosowanie teorii prawdopodobieństwa, co pozwala na jawne uwzględnienie niepewności właściwej procesowi wyszukiwania informacji tekstowej. Wadą modelu, w jego podstawowej postaci, jest uboga reprezentacja dokumentów i zapytań.

W kolejnych latach opracowano wiele wersji modelu probabilistycznego różniących się przede wszystkim sposobami szacowania prawdopodobieństw. Jedną z szerzej znanych implementacji modelu probabilistycznego opracowano w ramach projektu o nazwie OKAPI [191].

Wśród innych podejść opartych na teorii prawdopodobieństwa wymienić można modele odwołujące się do bayesowskich sieci wynikania (w skrócie $BN$) [210, 209]. Modele oparte na sieciach bayesowskich zbudowane są zazwyczaj z dwóch części: sieci dokumentów reprezentującej kolekcję dokumentów oraz sieci zapytań. Sieć dokumentów jest zasadniczo budowana tylko raz i nie ulega zmianom, natomiast sieć zapytań ulega zmianie podczas interakcji z użytkownikiem. Sieci wynikania pozwalają na „rozmycie" pojęć oraz tworzenie zależności między dokumentami wykraczających poza zakres słów kluczowych użytych w zapytaniu. Poszczególne rodzaje sieci mogą się różnić pomiędzy sobą właściwościami [141].

Omówione pokrótce klasyczne modele IR przedstawione zostały w ich oryginalnej, najprostszej postaci. Stały się one punktem wyjścia dla licznych rozszerzonych wersji. W dalszej części książki omawia się nowy model zaproponowany z zastosowaniem logiki rozmytej.

## 6.2 Wskaźniki efektywności systemów IR

Efektywność systemów wyszukiwania informacji tekstowej wyraża się przede wszystkim ich zdolnością do wybrania względem podanego zapytania wszystkich dokumentów relewantnych i tylko dokumentów relewantnych. Oceniana jest ona zazwyczaj z użyciem dwóch podstawowych wskaźników efektywności: *dokładności*, oznaczanej jako $P$ (ang. *Precision*), i *kompletności*, oznaczanej jako $R$ (ang. *Recall*).

W literaturze zaproponowano wiele innych wskaźników (por. np. [199, 5, 141]). Większość spośród nich odwołuje się do liczności dwóch podzbiorów dokumentów:

- relewantnych względem danego zapytania;
- wyszukanych w odpowiedzi na zapytanie.

**Dokładność ($P$)** Określa zdolność systemu do odrzucania dokumentów nierelewantnych. Jest to stosunek liczby wyszukanych dokumentów relewantnych do ogólnej liczby dokumentów wyszukanych.

$$P = \frac{|\{dokumenty\ relewantne\} \cap \{dokumenty\ wyszukane\}|}{|\{dokumenty\ wyszukane\}|} \tag{6.25}$$

**Kompletność ($R$)** Określa zdolność systemu do wybierania dokumentów relewantnych. Jest to stosunek liczby wyszukanych dokumentów relewantnych do całkowitej liczby dokumentów relewantnych względem zapytania.

$$R = \frac{|\{dokumenty\ relewantne\} \cap \{dokumenty\ wyszukane\}|}{|\{dokumenty\ relewantne\}|} \tag{6.26}$$

Użycie jednego z dwóch powyższych wskaźników z osobna nie daje rzetelnej oceny działania systemu wyszukiwania. Na przykład, zwracając w odpowiedzi na zapytanie wszystkie dokumenty z kolekcji system uzyskuje 100% wartość wskaźnika $R$. W celu przezwyciężenia tego problemu stosuje się różne kombinacje obydwu wskaźników. Popularne jest stosowanie ich średniej harmonicznej nazywanej wskaźnikiem $F$.

Wskażniki $P$ i $R$ zdefiniowane są terminach zbiorów dokumentów zwracanych w odpowiedzi na zapytanie. W praktycznych zastosowaniach (por. przeglądarki internetowe) system zwraca zazwyczaj listę dokumentów uporządkowaną według ich oszacowanej przez system relewantności

(według obliczonego stopnia dopasowania). W takim przypadku wartości wskaźników dokładności i kompletności oblicza się dla każdej pozycji $m$ tej listy z osobna, przyjmując na użytek wzorów (6.25)-(6.26), że zbiór *dokumentów wyszukanych* obejmuje pierwszych $m$ dokumentów z listy. Tak policzone wartości wskaźników najlepiej jest zilustrować na wykresie zależności *R-P*. Wykres *R-P* jest tworzony dla zapytania w ten sposób, że na osi $OX$ odkładane są wartości wskaźnika $R$, a na osi $OY$ wartości wskaźnika $P$. Analizujemy listę dokumentów zwróconą przez system, począwszy od pierwszej pozycji. Odnajdujemy kolejne pozycje na liście, na których występują dokumenty relewantne. Dla tych pozycji obliczamy wartości wskaźników $R$ oraz $P$ i zaznaczamy na wykresie punkty o tak obliczonych współrzędnych. Otrzymany wykres zależności *R-P* pokazuje wartość wskaźnika dokładności $P$ tylko dla niektórych wartości kompletności $R$ - dla tych wartości, które faktycznie realizują się dla danego uporządkowania. Poza tym dla jednej wartości $R$ mamy kilka wartości $P$. W celu usunięcia tych niedogodności na podstawie wykresu zależności $R - P$ tworzy się wykres interpolowanych wartości wskaźnika dokładności dla całego przedziału zmienności wartości wskaźnika kompletności, czyli dla całego przedziału [0,1]. Szczegóły mozna znaleźć na przykład w [5, 171].

Wykres zależności *R-P*, pokrewny znanym z klasyfikacji krzywym ROC, pozwala na szczegółową analizę zachowania się danego systemu wyszukiwania. Jednak często do porównania działania wielu systemów dogodniejszym może być zastosowanie syntetycznego wskaźnika jakości, wyrażonego jedną wartością liczbową. W literaturze zaproponowano wiele takich wskaźników. Wśród nich wymienić można:

**11-punktowa Interpolowana Średnia Dokładność** (ang. *11-point Interpolated Average Precision, 11-AVP)* obliczany jako średnia wartość interpolowanego wskaźnika dokładności dla 11 poziomów wartości wskaźnika kompletności równych *0.0, 0.1, 0.2,. . . , 1.0*;

**Average Precision at Seen Relevant Documents** obliczany jako średnia wartości wskaźnika dokładności policzonego dla każdej pozycji w uporządkowaniu wynikowym, na której występuje dokument relewantny;

**K-Dokładność** (ang. *R-Precision*) obliczany jako wartość wskaźnika dokładności dla $k$-tej pozycji listy dokumentów zwracanej przez system, przy czym $k$ jest równe liczbie dokumentów relewantnych

dla danego zapytania:

$$\frac{|\{\textit{liczba dokumentów relewantnych na pierwszych } k \textit{ pozycjach listy}\}|}{k} \tag{6.27}$$

**Znormalizowana Kompletność** (ang. *Normalized Recall*) oblicza się jako miarę podobieństwa wykresu wartości wskaźnika kompletności dla poszczególnych pozycji wynikowego uporządkowania dokumentów do wykresu odpowiadającego uporządkowaniu idealnemu, czyli takiemu, w którym wszystkie dokumenty relewantne występują na kolejnych pozycjach początkowych. Wartość wskaźnika jest odwrotnie proporcjonalna do wielkości obszaru, pomiędzy wykresem uzyskanym dla danego uporządkowania i wykresem dla uporządkowania idealnego;

**Znormalizowana Dokładność** (ang. *Normalized Precision*) oblicza się analogicznie do wskaźnika znormalizowanej kompletności, z tym że używane są wykresy wartości wskaźnika dokładności dla poszczególnych pozycji wynikowego uporządkowania dokumentów.

ISSN 0208-8029
ISBN 83-894-7551-0

INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK
tel.: (+48) 22 3810246 / 22 3810277 / 22 3810241 / 22 3810273
e-mail: biblioteka@ibspan.waw.pl