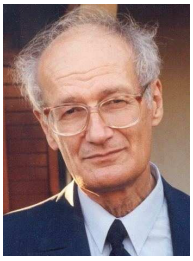
Bogdan Ionescu

Ionuț Mironică

Conceptul de Indexare Automată după Conținut în Contextul Datelor Multimedia

București, 2013

Prefaţă



Ce vrea să zică asta - indexarea după conţinut - cititorul va găsi în primul capitol, dar sunt tentat să zic şi aici, în aceste rânduri, câteva cuvinte: problema nu e chiar nouă. Cu ceva zeci de ani în urmă am aflat că pe alte meleaguri oamenii se ocupau, pentru cuvinte, cu alcătuirea unor asemenea dicţionare. Cele alfabetice, pe care le avem şi noi, îţi explică ce vrea să zică un cuvânt pe care îl ai dar al cărui sens nu îl ştii; dar sunt şi probleme de alt fel: acolo era un exemplu de întâmplare în academia spaniolă - un vorbitor nu-şi aducea aminte cum se cheamă un om născut pe vapor (noi n-avem cuvânt pentru acest concept). Ne trebuie dicţionare care să ne ducă de la concept la cuvânt. Despre unele popoare primitive se zice că aveau zeci de cuvinte pentru a denumi diferite tipuri de nori; noi n-avem, dar am putea eventual descrie forma lor, mişcarea lor, ca să precizăm la care ne referim când vrem să povestim o întâmplare concretă.

Într-o bibliotecă de un miliard de cărţi, cu câte 500 de pagini fiecare şi cu 2.000 de semne pe pagină avem nevoie doar de 50 de cifre binare pentru a identifica orice literă, ceea ce mi se pare extrem de puţin - la îndemâna umanului: le cuprindem cu ochiul dintr-o privire, pe un rând. Oare nu e posibil să avem căi/o cale de a ajunge la "obiectul" dorit dintr-o colecţie vastă, cunoscându-l prin calităţile sale (făcute cumva măsurabile: da-nu, roşu-albastru-galben-verde, o valoare întreagă între 1 şi 100, 17 grade de turtire a unui cerc în elipsă, etc.)? "Obiectele" de care vorbeam pot fi entităţi foarte complexe: o imagine, o secvenţă de film mut, entităţi "multimodale" (vorbă, sunete, imagini, text, etc.). Parcă suntem tentaţi a zice da. Dar acum vine partea dificilă a problemei, şi în acelaşi timp frumoasă prin efortul de

creaţie pe care ni-l cere (aspectul care ne provoacă, ne desfide, englezul ar zice "chalenging"): pe de o parte, în cazul concret al unei colecţii de un tip dat (de pietre, de gâze, de filme), care sunt atributele, cum le definim ca să caracterizăm cât mai compact şi mai corect, acea colecţie; pe de altă parte, în faţa unui obiect din colecţie, cum măsurăm *automat*, adică nu prin intervenţia omului (în cazul ăsta avem nevoie de un specialist în domeniu!), aceste atribute.

Fără acest mic amănunt aici, "automat", suntem pierduţi fiindcă operaţia manuală de adnotare cu atribute a obiectelor este consumatoare de timp în aşa măsură că ne face întreprinderea lipsită de sens.

În momentul de faţă al scurtei noastre istorii de câteva sute de ani, suntem în pericol de a fi "înecaţi în informaţii" care pe de o parte multe ne sunt vitale şi pe de alta, în ansamblul lor ne copleşesc, fără a putea ajunge la cele de care avem nevoie suntem ca însetatul din pustiu peste care năvăleşte marea. Indexarea automată după conţinut ne poate salva.

Extras din prefaţa cărţii "Analiza şi Prelucrarea Secvenţelor Video: Indexarea Automată după Conţinut", Editura Tehnică Bucureşti, 2009.

Prof. univ. dr. ing. Vasile BUZULOIU (1938 - 2012)
Bucureşti 17 Noiembrie 2008

Cuvântul autorului

Indexarea automată după conţinut a datelor este un domeniu de actualitate ce câştigă din ce în ce mai mult teren datorită necesităţii crescânde de exploatare a volumelor mari de date multimedia.

Progresul tehnologic al dispozitivelor de achiziţie şi prelucrare a datelor (terminale mobile, sisteme de calcul, medii de stocare, dispozitive de redare şi captură audio-video) cât şi a infrastructurii de transmisie de date (protocoale de transmisie fără fir: WiFi, Bluetooth, reţele LAN de mare viteză, telefonia multimedia 3G şi 4G) au condus practic la simplificarea stocării, transmisiunii şi prelucrării volumului important de date specific multimedia (video, imagini, sunet, text).

Mărturie în acest sens este răspândirea Internet-ului în tot mai multe medii sociale şi posibilitatea de accesare a acestuia de pe o categorie tot mai diversă de dispozitive electronice. La acestea se adaugă şi succesul imens de care se bucură reţelele de socializare şi platformele multimedia "on-line", Facebook, Twitter, Linkedln, Google+, YouTube, Dailymotion, Picasa, Flickr sunt doar câteva exemple dintre acestea.

Dinamica partajării datelor pe Internet este una copleşitoare, aceasta realizându-se practic "în timp real" de pe orice terminal multimedia. Următoarele statistici sunt edificatoare în acest sens: în 2012 mai mult de 72 de ore video sunt încărcate în fiecare minut pe platforma YouTube, mai mult de 500 de ani de video de pe platforma YouTube sunt vizualizaţi zilnic de pe platforma de socializare Facebook, mai mult de 700 de înregistrări video de pe YouTube sunt partajate în fiecare minut pe reteaua de socializare Twitter.

În societatea curentă, accesul la informaţia multimedia a devenit parte integrantă din viaţa noastră de zi cu zi. Problema cu care ne confruntăm nu este lipsa informaţiei, ci imposibilitatea de a selecta dintr-un vast amalgam

de date, informaţiile utile. Această problemă este cu atât mai dificilă cu cât conţinutul acestor date a devenit din ce în ce mai complex.

Până nu demult, când făceam referire la informaţie multimedia ne adresam imaginilor, înregistrărilor audio sau eventual video. În prezent conceptul de multimedia vine să reunească toate aceste informaţii sub umbrela unei singure paradigme şi anume aceea a reprezentării multimodale a informaţiei. Datele multimedia sunt practic "metadate" ce reunesc orice tip de informaţie video, audio şi textuală. Metodele de prelucrare trebuie să se adapteze acestor noi cerinţe în care analiza de conţinut este unitară şi nu realizată independent pentru fiecare sursă de informaţii.

În acest context, lucrarea de faţă vine să realizeze o trecere în revistă a domeniului indexării automate după conţinut a datelor multimedia şi să discute soluţiile existente.

Lucrarea este structurată în felul următor. În primul rând este introdusă problematica indexării datelor şi aplicaţiile acesteia (Capitolul 1). Mai departe, este prezentat detaliat mecanismul de funcţionare al unui sistem de indexare ce implică descrierea conţinutului datelor, mecanismul de căutare a datelor şi respectiv interacţia cu utilizatorul (Capitolul 2). Capitolul 3 realizează o trecere în revistă a tehnicilor de descriere a conţinutului datelor folosind informaţia vizuală, audio şi respectiv textuală. Capitolul 4 se interesează de tehnicile de fuziune a informaţiei specifice abordărilor multimodale ce exploatează date heterogene. Mai departe este adusă în discuţie problema evaluării similarităţii datelor (Capitolul 5). Tehnicile de interacţie cu utilizatorul în vederea îmbunătăţirii performanţelor de indexare sunt prezentate în Capitolul 6. Capitolul 7 discută problematica vizualizării informaţiei multimedia şi în special a datelor video. În final, Capitolul 8 prezintă o serie de modalităţii de evaluare subiectivă şi obiectivă a performanţelor unui sistem de indexare iar Capitolul 9 concluzionează lucrarea sintetizând paradigmele actuale ale sistemelor de indexare.

Lucrarea de faţă se doreşte a fi un studiu introductiv al domeniului, furnizând cititorului o vedere de ansamblu asupra tehnicilor de prelucrare aferente sistemelor de indexare şi a avantajelor şi limitărilor acestora. Pentru o descriere detaliată, cititorul este îndrumat să consulte referinţele bibliografice furnizate.

Ş.l. univ. dr. ing. Bogdan IONESCU
Bucureşti 26 Aprilie 2013

Cuprins

CAPITOLUL 1

Introducere

Dacă în urmă cu aproximativ un deceniu, cantitatea de informaţie multimedia disponibilă era una redusă, în zilele noastre putem vorbi despre o *explozie informaţională*. Accesul la informaţia multimedia sau "conţinut", fie că este vorba de imagini, sunet, text sau video, a devenit practic parte integrantă din viaţa noastră de zi cu zi. Evoluţia tehnologică a dispozitivelor de achiziţie şi prelucrare a datelor (terminale mobile, sisteme de calcul, medii de stocare, dispozitive de redare şi captură audio-video) cât şi a infrastructurii de transmisie de date (protocoale de transmisie fără fir: WiFi, Bluetooth, reţele LAN de mare viteză, telefonia multimedia 3G şi 4G) au dus la creşterea exponenţială a volumului multimedia prin facilitarea stocării şi prelucrării acestuia.

La acestea contribuie semnificativ şi răspândirea Internet-ului în tot mai multe medii sociale precum şi succesul imens de care se bucură reţelele de socializare "on-line" (exemplu: Facebook[1], Twitter[2], Linkedln[3], Google+[4]) cât şi platformele web multimedia (exemplu: YouTube[5], Dailymotion[6], Picasa[7], Flickr[8]). Pe lângă producţia de conţinut multimedia să spunem comercial

[1] `https://www.facebook.com`
[2] `https://twitter.com`
[3] `http://ro.linkedin.com`
[4] `https://plus.google.com`
[5] `https://www.youtube.com`
[6] `https://www.dailymotion.com`
[7] `http://picasa.google.com`
[8] `https://www.flickr.com`

(realizat de companii în vederea comercializării), accesul la reţele de socializare şi platfome web a condus practic la facilitarea posibilităţii de a partaja şi accesa date multimedia personale, generate de utilizatorii de rând, precum fotografii, filme din colecţiile personale, reportaje, "video blogging" şi aşa mai departe. Acestea reprezintă o sursă imensă de conţinut multimedia, să luam ca exemplu reţeaua de socializare Facebook care în 2012 însuma nu mai puţin de 1.2 miliarde de utilizatori ce partajează informaţii multimedia.

În prezent dinamica partajării datelor pe Internet este una copleşitoare aceasta realizându-se practic "în timp real" de pe orice terminal multimedia, atât mobil (de exemplu telefonul mobil) cât şi fix. Prin simpla apasare a unui buton, o înregistrare video sau imagine poate fi încărcată imediat "on-line". Următoarele statistici sunt edificatoare în acest sens: în 2012 mai mult de 72 de ore video sunt încărcate în fiecare minut pe platforma YouTube, mai mult de 500 de ani de video de pe platforma YouTube sunt vizualizaţi zilnic de pe platforma de socializare Facebook, mai mult de 700 de înregistrări video de pe YouTube sunt partajate în fiecare minut pe reteaua de socializare Twitter. Dintre informaţiile multimedia cel mai frecvent tranzacţionate, conţinutul video "on-line" reprezintă cea mai mare categorie de date vehiculate pe Internet, cuprinzând în 2012 26% din traficul total de date (sursa CISCO Systems[9]). Până în 2015 se estimează că mai mult de 1 million de minute video (674 zile) vor traversa Internetul în fiecare secundă.

Astfel că problema cu care ne confruntăm acum nu este lipsa de informaţie, ci, dimpotrivă imposibilitatea de a selecţiona din volumul informaţional imens disponibil, *informaţia utilă* căutată. Am ajuns în punctul în care acest lucru nu mai poate fi realizat de operatori umani şi este necesară preluarea acestei sarcini de către calculator.

Această problematică de cercetare se găseşte la afluenţa unor domenii precum prelucrarea şi analiza semnalelor ("signal processing"), vederii asistate de calculator ("computer vision") şi al clasificării datelor ("data mining"). Importanţa acestei direcţii de cercetare a dat naştere unor domenii dedicate precum "multimedia" şi al "căutării de informaţii" ("information retrieval"). Cercetările actuale vizează dezvoltarea de metode *automate* capabile să înţeleagă conţinutul datelor şi să il pună la dispoziţia utilizatorului într-un mod foarte apropiat de modul în care o persoană ar realiza acest lucru (apropiat de modul de percepţie uman).

O potenţială soluţie la problema căutării informaţiei multimedia a fost discutată cu mai mult timp în urmă în contextul căutării de imagini şi consta în folosirea de *tehnici de indexare automată după conţinut* [Smeulders 00]. Transpuse în contextul actual tehnologic, aceste tehnici trebuie acum să se

[9] `http://www.cisco.com`

adapteze, pe de-o parte unui volum imens de date, de exemplu în cazul video doar 1 minut este echivalentul a 1.500 de imagini statice şi astfel o singură secvenţă video echivalează conţinutul unei întregi colecţii de imagini; cât şi manipulării de conţinut temporal, în mişcare (video) şi multimodal (text-sunet-imagine). În ciuda unei disponibilităţii de putere de calcul în continuă creştere (în prezent un simplu telefon mobil foloseşte procesoare cu patru nuclee de prelucrare şi frecvenţe de 1.6 GHz) compexitatea acestei probleme necesită optimizarea şi paralelizarea metodelor. Acestea trebuie să fie eficiente computaţional pentru a putea fi aplicate la scară largă colecţiilor de pe Internet.

Cât de departe este tehnologia actuală pentru a realiza acest lucru? Să luam ca exemplu cazul simplificat al căutării după conţinut al imaginilor. În Figura 1.1 am prezentat rezultatele obţinute pentru căutarea unor imagini ce conţin "nuferi galbeni" folosind motorul de căutare propus de Google şi anume Google Search by Image[10] - considerat una dintre tehnologiile de vârf în prezent. Pentru a specifica datele dorite, am furnizat ca exemplu o imagine.

Figura 1.1: Exemplu de căutare după conţinut pentru o imagine cu un nufăr galben ("water lily", imagine stânga) folosind motorul de căutare Google Search by Image. Imaginile din dreapta reprezintă primele şase rezultate obţinute în ordinea descrescătoare a similarităţii (ordine sus în jos şi de la stânga la dreapta).

Se poate observa că în ciuda faptului că imaginile returnate au proprietăţi vizuale similare cu imaginea dată drept referinţă, semnificaţia semantică a acestora poate fi complet diferită. De exemplu, primim ca rezultat alte tipuri de flori sau chiar o persoană cu un tricou având culori similare. Cu toate că sistemele de căutare după conţinut a imaginilor au în acest moment aproape două decenii de existenţă, şi aici ne referim nu la tehnologia în sine ci la sisteme funcţionale, un exemplu în acest sens fiind sistemul Query By Image Content – QBIC propus de IBM în 1995 [Flickner 95], tehnologia actuală nu

[10] `http://images.google.com`

este încă capabilă să atingă un nivel apropiat de modul în care o persoană ar rezolva problema căutării, manual.

Tehnicile de căutare după conţinut a informaţiei video sunt şi mai puţin dezvoltate în acest moment limitându-se în principal în a fi extensii temporale ale celor aplicate în cazul imaginilor statice (de exemplu pentru a lua în calcul dimensiunea temporală de mişcare). În prezent, nu există un sistem de căutare după conţinut video disponibil public, încercările existente fiind doar experimentale, adaptate la baze video "off-line" de dimensiuni reduse (în cazul cel mai bun de sute de mii de secvenţe) şi limitate în a se adresa unor aplicaţii particulare (de exemplu căutarea de conţinut de ştiri, sport, catalogarea colecţiilor de filme după gen, identificarea conţinutului de animaţie şi aşa mai departe).

Platformele de căutare multimedia existente sunt limitate în a folosii doar informaţie textuală, precum descrierile asociate de către utilizatori datelor. De exemplu, o înregistrare cu turnul Eiffel poate fi însoţită de o descriere de genul "vizită turnul Eiffel, Paris 2013". Utilizatorul va căuta informaţia dorită furnizând tot o descriere textuală a acesteia, ca de exemplu caută toate înregistrările cu "turnul Eiffel", furnizând aceste cuvinte cheie. Informaţia furnizată va fi comparată cu cea asociată datelor obtinând ca rezultat secvenţele corespunzătoare, precum secvenţa etichetată anterior. Aparent problema pare a fi rezolvată. Totuşi, informaţia textuală este limitată în a furniza doar o descriere globală şi parţială a conţinutului. În exemplul anterior, sistemul pe baza descrierilor existente nu va fi capabil de exemplu să identifice prezenta unei anumite persoane în acea înregistrare deoarece această informaţie lipseşte din descriere. Mai mult, descrierile textuale nu pot fi determinate în mod automat, necesitând intervenţia umană. Extrapolând această problemă la dimensiunea bazelor multimedia de pe Internet, asocierea de descrieri textuale care să detalieze conţinutul datelor video devine practic imposibilă.

Soluţia la problema căutării după conţinut a datelor multimedia nu se găseşte la nivel de modalitate individuală şi anume la nivel de imagine, video, sunet sau chiar text. Soluţia ţine de o abordare globală interdisciplinară a acestei problematici prin interacţionarea informaţiilor multimodale extrase din toate sursele de informaţie disponibile, de la culoare, textură, forme, mişcare, informaţie temporală pănă la sunet, voce, text şi aşa mai departe. Aceasta constituie de fapt tendiţa actuală de cercetare. Folosirea independentă a surselor de informaţie se dovedeşte ineficientă pentru a rezolva o problemă atât de complexă precum înţelegerea automată a conţinutului datelor multimedia. Ca referinţă în acest sens sunt campaniile TRECVID Video

Retrieval Evaluation Benchmarking Campaign[11], MediaEval Benchmarking Initiative for Multimedia Evaluation[12], ImageCLEF The CLEF Cross Language Image Retrieval Track[13] sau PASCAL Challenge - Pattern Analysis, Statistical Modelling and Computational Learning[14] ce anual prezintă tehnologiile şi bunele practici curente din domeniu. Cititorul se poate raporta la acestea pentru o vedere de ansamblu a progresului tehnologic actual în acest domeniu.

În cele ce urmează vom face o trecere în revistă a tehnicilor ce stau la baza procesului de indexare după conţinut, a tehnicilor de descriere a conţinutului datelor şi a surselor informaţionale exploatate, a tehnicilor de fuziune a informaţiilor multimodale, tehnicilor de integrare a opiniei utilizatorului în procesul de indexare, a problematicii vizualizării conţinutului multimedia, a modului de evaluare al performanţelor unui sistem de indexare încheind cu prezentarea barierelor actuale ale sistemelor de indexare după conţinut.

[11] `http://trecvid.nist.gov`
[12] `http://www.multimediaeval.org`
[13] `http://www.imageclef.org`
[14] `http://pascallin.ecs.soton.ac.uk/challenges/VOC`

CAPITOLUL 2

Mecanismul de indexare după conținut

Conceptul de indexare folosit pentru căutarea datelor este definit ca fiind *procesul de adnotare* a informaţiei existente într-o colecţie de date, prin adăugarea de informaţii suplimentare relative la conţinutul acesteia, informaţii numite şi *indici* de conţinut [Kyungpook 06]. Această etapă este necesară accesării colecţiei de date, deoarece permite catalogarea automată în funcţie de conţinut a datelor.

Într-o colecţie de date suficient de vastă, putem spune că datele care nu au fost adnotate sunt practic inexistente pentru utilizator. Un exemplu simplu de sistem de indexare este însuşi sistemul de fişiere al oricărui calculator personal. Acesta ne furnizează datele aflate pe diversele medii de stocare (disc dur, memorie externă, etc.) sub formă de fişiere ce sunt indexate după informaţii precum nume, extensie, dată, şi aşa mai departe. Să ne imaginăm situaţia în care un fişier a fost omis din această listă de indici, cu toate că el este prezent fizic pe suportul de stocare, acesta va fi invizibil şi inaccesibil pentru utilizatorul de rând.

Procesul de adnotare a datelor este văzut din două perspective: pe de-o parte există *adnotarea manuală*, iar pe de altă parte *adnotarea automată*. Gradul de complexitate al adnotării este direct proporţional cu nivelul de detaliu dorit pentru accesarea datelor. Dacă se doreşte ca utilizatorul să poată accesa datele folosind criterii mai complexe, ca de exemplu căutarea unei anumite secvenţe video pentru care nu se cunoaşte nici numele, nici extensia fişierului, dar totuşi utilizatorul dispune de informaţii referitoare la conţinutul vizual al acesteia, în această situaţie, procesul de indexare va fi mult mai complex, necesitând întelegerea de către calculator a conţinutului

datelor.

Astfel, în cazul unei indexări după criterii complexe de conţinut, adnotarea manuală este foarte dificil de realizat, deoarece necesită un număr important de operatori umani. Aceştia ar trebui să "răsfoiască" manual întregul conţinut al bazei de date pentru definirea indicilor de conţinut. Luând în calcul faptul că o astfel de colecţie de date este în prezent practic nelimitată (exemplul sunt colecţiile de pe Internet), indexarea manuală devine imposibilă. În acest moment, cercetările existente în domeniu se focalizează pe dezvoltarea de algoritmi de adnotare automată a conţinutului, mai ales în cazul datelor ce necesită un timp important pentru vizualizare, ca de exemplu documentele video.

Cu toate că adnotarea conţinutului datelor este soluţia optimală pentru a accesa informaţia utilă dintr-o vastă colecţie de date, aceasta nu este şi suficientă. Adnotarea în sine nu oferă decat o serie de date suplimentare, putem spune, de nivel semantic inferior ("low-level"), care deseori sunt inaccesibile utilizatorului neavizat. Pentru a accesa baza de date, utilizatorul trebuie să dispună de o modalitate prin care să poată accesa sau vizualiza uşor datele, fie pe baza indicilor, fie în mod direct. Aceasta trebuie să aibă o funcţionalitate naturală şi intuitivă. Sistemul care permite utilizatorului să vizualizeze conţinutul bazei de date poartă numele de *sistem de navigare*.

Pe de altă parte, accesul la date presupune un proces de căutare. Utilizatorul trebuie să mai dispună, pe lângă sistemul de navigare, de un mecanism care să-i permită căutarea informaţiilor dorite în baza de date. Căutarea se realizează prin formularea de cereri de căutare sau "queries". Pentru uşurinţă, o astfel de cerere trebuie să fie exprimată într-un limbaj natural, apropiat de limbajul uman, cum ar fi de exemplu "caută filmele de acţiune" sau "caută imaginile ce conţin peisaje". Sistemul care răspunde acestor cerinţe poartă numele de *sistem de căutare*. Figura 2.1 sintetizează aceste aspecte prezentând schematic modul de funcţionare al unui sistem generic de indexare a datelor.

Astfel, pentru a sintetiza, mecanismul de indexare şi căutare a datelor presupune realizarea următoarelor etape:

- **descrierea conţinutului datelor**: într-o primă etapă, informaţia propriu-zisă din baza de date este reprezentată prin intermediul atributelor de conţinut, informaţii pe baza cărora se realizează întregul proces de indexare (vezi Secţiunea 2.1);

- **formularea cererii de căutare**: utilizatorul furnizează o descriere a datelor pe care doreşte să le găsească prin formularea unei cereri de căutare ("query"). Acest lucru poate fi realizat folosind un exemplu

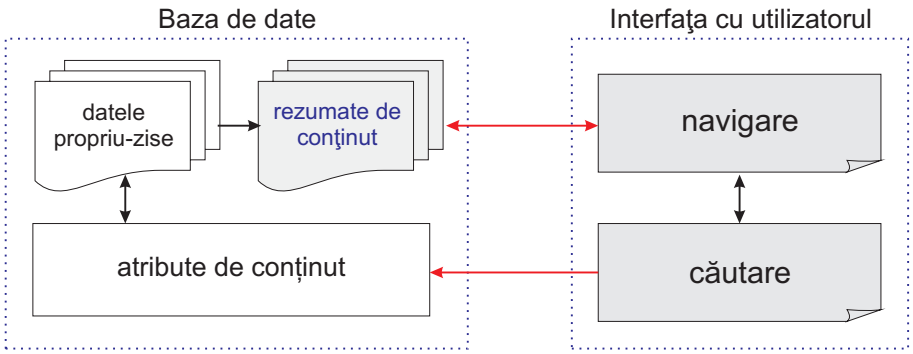


Figura 2.1: Principiul de funcţionare al unui sistem de indexare după conţinut.

a ceea ce caută, folosind o descriere textuală a conţinutului datelor căutate, pe baza unei descrieri grafice schematice a proprietăţilor datelor căutate şi aşa mai departe (vezi Secţiunea 2.2);

- **conversia în descriptori**: sistemul de căutare traduce cererea utilizatorului în atribute de conţinut folosind un mecanism similar cu cel folosit la adnotarea conţinutului bazei de date. Aceşti descriptori pot fi proprietăţi de culoare, forme, informaţie audio sau de mişcare (vezi Secţiunea 2.1);

- **căutarea propriu-zisă**: căutarea se realizează prin compararea atributelor cererii de căutare cu cele deja stocate în baza de date. Folosind diverse măsuri de distanţă şi similaritate între atribute, sistemul va căuta datele ce sunt cele mai apropiate (similare) de criterile formulate (vezi Secţiunea 2.3);

- **interacţia cu utilizatorul**: rezultatele căutării sunt furnizate utilizatorului de regulă folosind sistemul de navigare. Acesta presupune o interfaţă vizuală intuitivă în care utilizatorul poate vizualiza eficient conţinutul datelor. În mod opţional, sistemul poate interacţiona cu utilizatorul ("feedback") pentru a îmbunătăţii performanţele sistemului, de exemplu înregistrând opinia utilizatorului cu privire la relevanţa datelor returnate de sistem (vezi Secţiunea 2.4).

În cele ce urmează vom detalia fiecare dintre aceste etape.

2.1 Descrierea conţinutului datelor

Într-o primă etapă, informaţia propriu-zisă din baza de date este reprezentată prin intermediul atributelor de conţinut. Sistemul va genera pentru fiecare document o colecţie de atribute ce vor caracteriza proprietăţile relevante ale conţinutul acestuia (denumiţi şi *descriptori*). De exemplu, documentul X poate fi descris de atributele A_1, A_2, ..., A_n unde valorile $\{a_1, a_2, ..., a_n\}$ formează descriptorul de conţinut. Atributele definesc ceea ce numim *spaţiul de caracteristici* al datelor, de regulă un spaţiu n-dimensional.

Atributele pot fi, fie date de *nivel semantic scăzut*, precum măsuri statistice, parametri numerici (de exemplu: histograme de culoare[1], câmpuri vectoriale de mişcare, histograme de orientare a contururilor din imagine), fie date simbolice de *nivel semantic superior* (de exemplu: nume obiecte de interes, percepţia culorilor, recunoaştere text "încrustat" în imagine, identificarea prezenţei umane). Cu alte cuvinte, informaţia iniţială heterogenă şi multimodală a fost convertită la o reprezentare uniformă într-un sistem unitar normalizat definit de spaţiul de caracteristici. Fiecare document va fi caracterizat astfel de o anumită valoare a acestor atribute, definind un punct unic în spaţiu.

Pentru a ilustra aceste aspecte, în Figura 2.2 am prezentat un exemplu concret de reprezentare a conţinutului în cazul înregistrărilor audio (şi în particular al sunetelor animalelor). Spaţiul de caracteristici este definit în acest caz de trei atribute şi anume: entropia Wiener[2] (A_1), amplitudine (A_2) şi continuitate în timp (A_3) (spaţiu tridimensional). Astfel, fiecare punct din spaţiu, P_i (reprezentat grafic de un cerc) cu $i = 1, ..., N$ unde N reprezintă numărul de înregistrări disponibile, reprezintă o înregistrare audio al cărei conţinut a fost descris de valorile atributelor A_1, A_2, A_3, şi anume $P_i = \{a_{i1}, a_{i2}, a_{i3}\}$ (vezi şi Secţiune 4 relativă la fuziunea descriptorilor). Dacă atributele sunt suficient de discriminatorii, înregistrările audio similare din punct de vedere al conţinutului trebuie să conducă la puncte apropiate spaţial (vezi cercurile de aceeaşi culoare) în timp ce înregistrările diferite trebuie să conducă la puncte distanţate spaţial (vezi punctele de culori diferite).

Tot în această etapă a descrierii conţinutului datelor, opţional, se pot

[1] histograma unei imagini este o măsură a probabilităţilor discrete de apariţie a culorilor (sau a intervalelor de culoare denumite şi bini) în imagine, valorile acesteia reprezentând numărul de apariţii al unei culori raportat la numărul total de pixeli. Astfel, histograma are sens de densitate de probabilitate a variabilei aleatoare determinată de valoarea unui pixel.

[2] entropia Wiener este definită ca fiind o măsură a lăţimii şi uniformităţii spectrului de putere audio. Ca referinţă, pe o scală de la 0 la 1, zgomotul alb (semnal aleator cu densitate spectrală de putere constantă) are o entropie 1 iar un ton pur are o entropie 0.

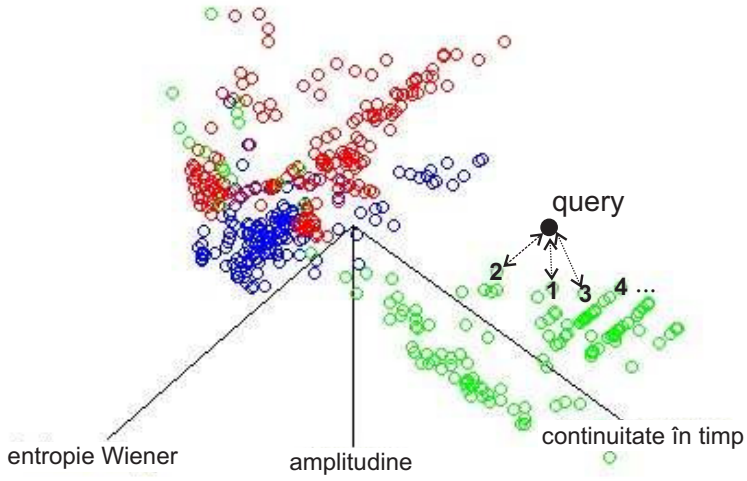


Figura 2.2: Exemplu de spaţiu de caracteristici în cazul conţinutului audio (sursă imagine programul de prelucrare audio "Sound Analysis Pro", http://soundanalysispro.com/).

genera *descrieri compacte*, precum scurte rezumate pentru secvenţele video sau pasaje de text reprezentative pentru documentele textuale. Rolul acestor rezumate este acela de a eficientiza vizualizarea conţinutului datelor. De exemplu, pentru o bază de documente video este practic imposibil ca utilizatorul să poată vizualiza rapid conţinutul acesteia. În acest caz, sistemul poate furniza utilizatorului doar câteva imagini reprezentative sau un rezumat de câteva secunde (exemplu un "trailer") ce redă informaţia cheie din secvenţă.

Dacă în urmă cu câţiva ani de zile extragerea de atribute putea fi considerată ca o etapă ce poate fi realizată "off-line", timpul de prelucrare nefiind critic, în prezent datorită dinamicii colecţiilor multimedia (să luăm ca exemplu YouTube ce raporta în 2012 o rată de încărcare de 72 de ore video pe minut) aceasta trebuie realizată mult mai rapid decât o prelucrare în timp real şi trebuie să poată fi scalabilă (să poată fi aplicată unor colecţii de date dinamice).

În acest punct al indexării, problema care apare este aceea a *relevanţei atributelor folosite*. Diversitatea surselor de informaţie disponibile face dificilă eficientizarea reprezentării datelor. Cu cât creşte dimensiunea spaţiului de caracteristici şi astfel numărul de atribute folosite la reprezentarea datelor cu atât tinde să crească redundaţa informaţiei şi să scadă puterea dis-

criminatorie a acestora. Un descriptor eficient este acela care maximizează informaţia reprezentată şi minimizează dimensionalitatea datelor. Mai multe informaţii relative la tehnicile existente de adnotare a conţinutului sunt prezentate în Secţiunea 3.

2.2 Formularea căutării

Sistemul de căutare va permite utilizatorului să localizeze informaţiile dorite pe baza formulării unei cereri de căutare, denumită şi "query" (concept similar celui utilizat în contextul bazelor de date numit şi interogare). În mod ideal, sistemul trebuie să poată permite ca aceasta să fie formulată într-un mod cât mai natural şi cât mai apropiat de modul de percepţie uman, pentru a putea fi la îndemâna oricărui utilizator.

Precizia rezultatelor căutării este în primul rând dependentă de modul de formulare a cererii de căutare a datelor sau cu alte cuvinte a modului de descriere a datelor care se doresc a fi găsite. Formularea adecvată a criteriilor de căutare nu este dependentă numai de sistemul de indexare aceasta depinzând în mare parte şi de utilizator.

În primul rând, nivelul de cunoaştere de către utilizator a caracteristicilor datelor căutate este primul factor ce influenţează căutarea. Se întâlnesc de regulă următoarele situaţii posibile [Maillet 03]:

- utilizatorul ştie cu siguranţă că datele căutate se află în baza de date. În acest caz, ţinta este unică iar utilizatorul va fi capabil să formuleze eficient cererea de căutare. Utilizatorul va repeta căutarea până când va obţine datele dorite;

- utilizatorul caută o anumită informaţie dar nu este sigur că aceasta este prezentă în baza de date. În acest caz, sistemul de indexare are rolul de a furniza algoritmi de căutare precişi şi eficienţi pentru ca utilizatorul să se decidă rapid dacă datele dorite sunt cu adevărat prezente în baza de date. Rafinarea ulterioară a căutării va permite identificarea mai precisă a datelor căutate;

- utilizatorul are informaţii vagi cu privire la ceea ce doreşte să găsească în baza de date. În această situaţie, sistemul de navigare poate fi folosit pentru "răsfoirea" preliminară a conţinutului şi identificarea unor informaţii de interes repoziţionând utilizatorul în una dintre cele două situaţii enumerate anterior.

Odată identificată informaţia dorită este necesar un formalism care să permită enunţarea cererii de căutare în sistemul de căutare. Acesta face

practic legătura dintre modul de percepţie uman şi reprezentarea informaţiei în sistemul respectiv. În funcţie de natura datelor căutate, în literatură există o multitudine de abordări posibile:

- **folosirea vorbirii**: în cazul căutării textuale (informaţie sub formă de text) se poate folosi direct comanda vocală. Utilizatorul vorbeşte practic ceea ce doreşte să caute, de exemplu: "caută prognoza meteo pentru astăzi" sau "caută informaţii despre posibilităţi de cazare în Paris". Comanda este transformată folosind algoritmi de recunoaştere automată a vorbirii în text care este comparat mai departe cu datele din bază. Datorită limitărilor tehnologice a sistemelor de indexare multimedia, o astfel de abordare foarte generală rămâne viabilă doar în cazul căutării de text, ca de exemplu pe Internet (vezi sistemul Siri de pe dispozitivele iPhone[3] sau sistemul Google Voice Search de pe dispozitivele cu sistem Android[4]);

- **folosirea de cuvinte cheie**: reprezintă o variantă intermediară a cazului anterior. Cererea de căutare este tot textuală dar este exprimată într-un mod mai restrictiv pe baza unor cuvinte cheie. Pentru ca acest mecanism să funcţioneze, datele căutate trebuie să aibă asociate descrieri textuale similare, descrieri ce sunt generate de regulă de utilizatori (de exemplu în momentul în care datele sunt încărcate pe o platformă media on-line) sau în mod automat (metodele de adnotare textuală automată a conţinutului multimedia - "tagging" - sunt totuşi încă destul de imprecise);

- **folosirea unui concept**: este de asemenea legată de specificarea unor cuvinte cheie. Diferenţa faţă de cazul anterior este dată de faptul că un concept este o noţiune destul de generală care face referire la o clasă de date şi nu neapărat la un obiect particular. De exemplu, se doreşte localizarea tuturor imaginilor ce conţin arbori, unde conceptul căutat este "arbore", sau a secvenţelor în care apar case, conceptul căutat fiind acela de "casă". Noţiunea de căutare de concepte este asociată în prezent datelor video şi constituie un pas intermediar în atingerea unui nivel de descriere textuală. La ora actuală sistemele de căutare după conţinut video sunt limitate în a fi antrenate la a răspunde unui număr destul de limitat de concepte (de ordinul miilor - vezi campania TRECVID[5]);

[3] `http://www.apple.com/ios/siri`
[4] `http://www.google.com/mobile/voice-search`
[5] `http://trecvid.nist.gov`

- **folosirea unui exemplu**: în acest caz, cererea este formulată folosind un model al datelor. De exemplu, utilizatorul caută toate imaginile asemănătoare cu o anumită imagine de care dispune, imaginea fiind furnizată ca exemplu (vezi sistemul de căutare Google Image Search[6]). Tot în această categorie intră şi cazul în care utilizatorul furnizează o descriere schematică a datelor căutate. De exemplu acesta nu dispune de o imagine de referinţă dar poate reprezenta schematic conţinutul dorit generând o schiţă a imaginii (poziţionarea anumitor categorii de obiecte, prezenţa anumitor culori şi aşa mai departe - vezi sistemul QBIC al Hermitage Museum[7]);

- **folosirea gesturilor**: un mod interesant de formulare a cererii de căutare o reprezintă gesticularea obiectului care se doreşte a fi căutat. Acest mod de căutare are totuşi un interes mai mult ştiinţific deoarece limitările fiziologice fac imposibilă reprezentarea oricărui obiect prin intermediul gesturilor (vezi un exemplu în [Shirahama 11]);

- **fredonarea unui pasaj audio**: în cazul căutării înregistrărilor audio, de regulă muzicale, o modalitate inedită de formulare a cererii de căutare constă în fredonarea unui pasaj din melodia dorită (vezi de exemplu sistemul Midomi[8]).

2.3 Căutarea datelor

Pentru a fi înţelese de sistem, cererile de căutare trebuiesc mai întâi convertite în atribute de conţinut folosind acelaşi mecanism ca şi în cazul adnotării iniţiale a bazei de date. În acest fel, cererea de căutare este reprezentată practic în spaţiul de caracteristici definit în etapa anterioară, prin intermediul unui descriptor. Mai departe, căutarea propriu-zisă se efectuează prin compararea valorilor acestui descriptor cu valorile descriptorilor datelor din bază.

Rezultatele căutării vor fi acele date ale căror valori sunt cele mai apropiate din punct de vedere al unuia sau a mai multor *criterii de similaritate*, de exemplu valorile minime ale unei mărimi de distanţă, folosirea unei baze de reguli de decizie şi aşa mai departe (vezi Secţiunea 5).

De exemplu, în cazul sistemului din Figura 2.2, cererea de căutare poate consta într-un exemplu de înregistrare audio. Utilizatorul doreşte localizarea

[6] `http://images.google.com`

[7] `http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English`

[8] `http://www.midomi.com`

tuturor înregistrărilor audio similare cu aceasta. Exemplul este convertit de sistem într-o serie de valori ale atributelor folosite la indexare, a_1, a_2, a_3, definind descriptorul de căutare: query=$\{a_{q1}, a_{q2}, a_{q3}\}$. Rezultatele căutării vor fi acele înregistări audio ce corespund punctelor cele mai apropiate de punctul definit de descriptorul de căutare (vezi Figura 2.2). Datorită subiectivităţii procesului de căutare, sistemul nu se limitează în a furniza un singur rezultat, ci va returna o clasificare ("ranking") a datelor în ordinea descrescătoare a similarităţii: poziţia 1 - data cea mai similară, poziţia 2 - următoarea dată cea mai similară, poziţia 3, şi aşa mai departe.

În acest punct al procesului de indexare, problema principală este definirea conceptului de *similaritate* dintre date. Dacă în cazul datelor numerice soluţia la această problemă se găseşte în matematică (prin conceptul de metrică), lucrurile nu sunt aşa de evidente în cazul datelor multimedia ce implică folosirea de descriptori de natură diferită (text-audio-vizuali). De exemplu, când două secvenţe pot fi considerate similare? sau două pasaje de text? Aceasta este o problemă subiectivă chiar şi pentru utilizator. Întreg procesul de indexare depinde de modul de definire al măsurii de distanţă, schimbarea acesteia poate conduce la rezultate complet diferite. O prezentare detaliată a măsurilor de distanţă folosite în contextul indexării datelor este realizată în Secţiunea 5.

2.4 Interacţia cu utilizatorul

Ultima etapă a procesului de indexare constă în interacţia cu utilizatorul. Aceasta este realizată de regulă prin intermediul *sistemului de navigare*. Sistemul de navigare este practic o interfaţă grafică ce deserveşte mai multe funcţionalităţi.

O primă funcţionalitate, independentă de procesul de căutare, este aceea de a furniza utilizatorului access direct la datele din bază. În funcţie de tipul datelor, poate fi necesară adoptarea unei strategii complexe. De exemplu, o bază de imagini poate fi vizualizată doar prin reprezentarea în miniatură a acestor imagini (folosind "thumbnails"). În cazul unei baze de secvenţe video, acest lucru poate fi realizat prin prezentarea a câtorva imagini reprezentative pentru fiecare secvenţă. Totuşi acest mod de prezentare nu este mereu suficient deoarece nu furnizează nici o informaţie relativă la conţinutul de mişcare (de acţiune) specific. O serie de soluţii de reprezentare a conţinutului video sunt discutate în Secţiunea 7.

O a doua funcţionalitate, şi poate cea mai importantă, este aceea de a pune la dispoziţia utilizatorului rezultatele obţinute în urma etapei descrise anterior şi anume a căutării după anumite criterii. Rezultatele sunt de re-

gulă vizualizate în ordinea descrescătoare a relevanţei (similarităţii) faţă de cererea de căutare.

În final, o altă funcţionalitate o constituie interacţia cu utilizatorul. În ciuda progresului actual al tehnicilor de descriere a conţinutului multimodal, procesul de indexare este inerent limitat de însăşi natura datelor (vezi Secţiunea 9). Trebuie să ţinem cont că practic imaginile şi înregistrările sunt de fapt nişte proiecţii limitate, bidimensionale, ale lumii înconjurătoare. Astfel, după cum am prezentat şi în introducerea acestui capitol, datorită puterii discriminatorii limitate a descriptorilor, rezultatele căutării nu sunt întotdeauna adaptate necesităţii utilizatorului. Pentru a ameliora acest aspect, de-a lungul timpului au fost studiate o serie de abordării ce tind să includă în procesul de indexare expertiza umană. Printre acestea, cea mai cunoscută poartă numele de "Relevance Feedback" (RF).

Un scenariu clasic de RF poate fi formulat în felul următor: pentru o anumită cerere de căutare rezultatele obţinute sunt puse la dispoziţia utilizatorului în ordinea descrescătoare a relevanţei. Mai departe, utilizatorul este solicitat să marcheze un număr limitat dintre acestea (de regulă de ordinul zecilor) în funcţie de relevanţa lor. Utilizatorul va marca datele ca fiind relevante - datele corespund perfect cererii de căutare sau nerelevante - datele nu corespund. Pe baza acestor informaţii, sistemul de căutare calculează o nouă reprezentare a datelor căutate şi returnează o rafinare a rezultatelor iniţiale. Cu alte cuvinte, acest proces îmbunătăţeşte răspunsul sistemului folosind informaţia de la utilizator pe post de "realitate" (sau "ground truth"[9]). Mai multe informaţii sunt prezentate în Secţiunea 6.

În acest punct al procesului de indexare avem la dispoziţie un lanţ complet de căutare ce porneşte de la definirea cererii de căutare şi se finalizează cu interacţia cu utilizatorul relativ la rezultatele obţinute. Problema care apare în acest punct este *evaluarea performanţei sistemului*. Cum putem evalua performaţele unui sistem de indexare? Faptul că acesta furnizează rezultate bune pentru o serie de cereri de căutare (vezi exemplul din Figura 6.1) îi garantează performanţa?

În realitate, performanţele sistemului variază în mod evident de la o cerere de căutare la alta (este posibil să avem date care sunt mai uşor de localizat datorită conţinutului acestora). Avem nevoie de o modalitate generală care

[9]termenul de "ground truth" îşi are originea în domeniul cartografiei şi implică procesul de colectare de informaţii despre un anumit fenomen, prin observarea practică pe teren a acestuia. Datele obţinute constituie "realitatea de teren" folosită pentru calibrarea, validarea şi interpretarea observaţiilor sau a măsurătorilor de la distanţă a fenomenului în cauză sau a altor fenomene similare. În contextul indexării, "ground truth" reprezintă datele pentru care se cunoaşte conţinutul acestora, de exemplu faptul că o imagine reprezintă un anumit obiect sau că o secvenţă video este de un anumit gen.

să evalueze performanţa sistemului, global, în orice situaţie. Acest lucru este realizat de regulă testând răspunsul acestuia la căutarea fiecărui document din baza de date considerată. Practic, fiecare document devine cerere de căutare.

Evaluarea performanţei rezultatelor este mai departe realizată fie subiectiv, de exemplu pe baza opiniei utilizatorilor, fie obiectiv folosind măsuri numerice de performanţă (exemplu numărul mediu de rezultate corecte, numărul mediu de rezultate eronate şi aşa mai departe). O trecere în revistă a abordărilor cel mai frecvent folosite în literatura de specialitate este prezentată în Secţiunea 8.

CAPITOLUL 3

Descrierea conţinutului multimodal

După cum am menţionat în secţiunile anterioare, procesul de adnotare al conţinutului datelor constă în *crearea atributelor* sau a descriptorilor de conţinut ce constituie baza sistemului de indexare. Practic căutarea datelor se realizează prin compararea valorilor acestor descriptori pentru cererea de căutare ("query") cu descriptorii informaţiilor existente în baza de date.

În acestă secţiune vom face o trecere în revistă a tehnicilor existente şi a surselor de informaţie folosite în cazul descrierii conţinutului multimedia urmând ca acestea să fie detaliate în secţiunile următoare. În principal putem identifica trei surse majore de informaţie, şi anume (vezi Figura 3.1):

- **informaţia vizuală**: acesta se referă la datele ce sunt percepute vizual, ca de exemplu culoare, formă, textură, mişcare, precum şi derivate din acestea;
- **informaţia audio**: se referă la datele ce sunt percepute sub formă de semnale sonore, ca de exemplu voce, vorbire, muzică, sunete ambientale sau zgomot;
- **informaţia textuală**: se referă la datele reprezentate sub formă de text (caractere) ce pot provenii din textul ataşat datelor (de exemplu textul ce înconjoară un obiect multimedia pe o pagina de web), textul obţinut prin recunoaşterea caracterelor ce apar "încrustate" în imagine (exemplu subtitrări), sau textul obţinut prin recunoaşterea vorbirii din informaţia audio.

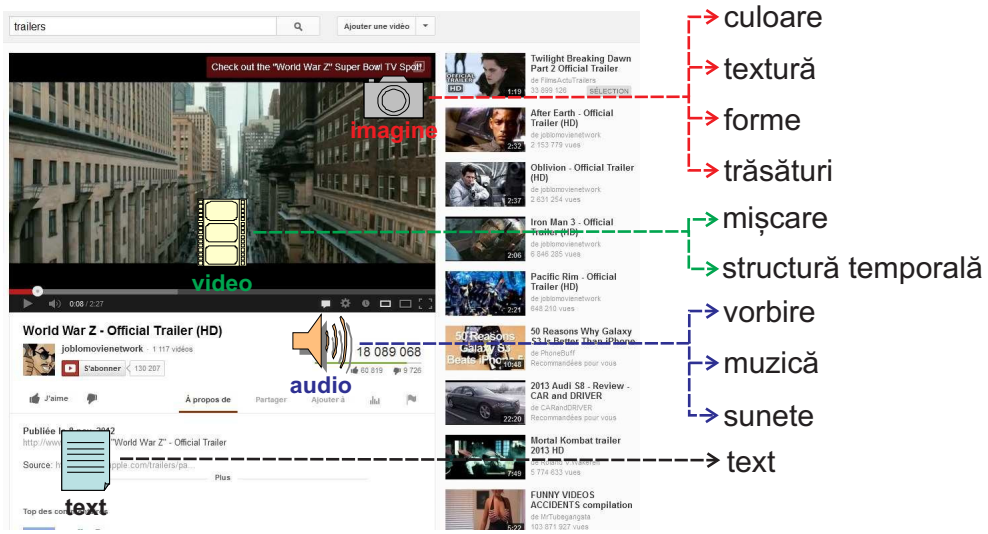


Figura 3.1: Surse de informaţie multimedia (sursă imagine platformă YouTube, http://www.youtube.com).

3.1 Informaţia vizuală

Informaţia de culoare reprezintă una dintre sursele de informaţie cel mai frecvent folosite în cazul descrierii conţinutului imaginilor. Acest lucru se datorează în principal faptului că însuşi sistemul vizual uman este bazat pe prelucrarea informaţiei de culoare (unde luminoase de diverse frecvenţe). Conţinutul de culoare este analizat pe baza reprezentării acestuia folosind un anumit model de reprezentare a culorilor[1] sau spaţiu de culoare.

Spaţiile de culoare folosite variază de la cele clasice, precum sistemul RGB (Red - Roşu, Green - Verde, Blue - Albastru), sisteme ce separă componenta de intensitate de componentele cromatice, precum YCbCr (Y - luminozitate, Cb, Cr - diferenţe cromatice), până la sisteme perceptuale în care culorile sunt structurate în aşa fel încât să reflecte modul de percepţie vizuală umană (culorile similare perceptual sunt alăturate în timp ce culorile opuse se găsesc separate), precum sistemul HSV (Hue - nuanţă, Saturation - saturaţie, Value - măsură a intensităţii), L*a*b* (L - luminozitate, a,b - diferenţe cromatice) în care distanţa perceptuală dintre culori tinde să fie proporţională cu distanţa matematică, sau HMMD (Hue - nuanţă, Max - măsură a gradului de

[1] un model de reprezentare a culorilor reprezintă un model matematic abstract ce descrie o culoare ca o combinaţie de numere, de regulă 3 sau 4 valori, ce corespund unor componente de culori primare (culorile primare sunt culori ce nu pot fi obţinute din combinaţia altor culori).

întunecare sau "shade", Min - măsură a gradului de luminare sau "tint", D - măsură a tonalităţii sau "tone") sistem ce oferă o serie de avantaje în contextul indexării după conţinut precum discretizarea mai eficientă a culorilor. Un studiu detaliat al spaţiilor de culoare este prezentat în [Trémeau 04].

O etapă premergătoare descrierii conţinutului de culoare constă în reducerea paletei de culoare[2] [Orchard 91]. Să luăm exemplul spaţiului RGB în care fiecare componentă de culoare este reprezentată pe 8 biţi ceea ce conduce la un număr total de 16.777.216 de culori posibile. În practică gestionarea unui număr semnificativ de culori este atât ineficientă deoarece ochiul uman nu este sensibil la micile variaţii de culoare, cât şi nerentabilă din punct de vedere computaţional. Paleta de culoare este redusă la un număr semificativ mai mic, de ordinul sutelor (exemplu de palete fixe: 256 de culori pentru paleta Windows pe 8 biţi[3]; sau 216 culori pentru paleta Webmaster[4]) folosind tehnici de cuantizare a culorilor. De asemenea, analiza conţinutului de culoare se poate realiza în urma segmentării imaginii în obiecte, proces de izolare a regiunilor din imagine ce corespund elementelor constituente ale scenei. În acest fel descrierea culorilor este realizată la nivel de obiect şi nu global la nivel de imagine.

Descrierea conţinutului de culoare se realizează de regulă folosind descriptori de nivel semnatic inferior precum histograme de culoare calculate în diverse spaţii de culoare, histograme ponderate, culori predominante, varianţa de culoare, parametri de intesitate, descrierea repartiţiei spaţială a culorilor, cât şi descriptori semantici precum prezenţa culorii pielii ("skin detection") ce indică prezenţa umană în scenă sau identificarea denumirii culorilor (asocierea de nume culorilor oferă informaţii asupra percepţiei acestora în imagine). Un studiu detaliat este prezentat în [Smeulders 00].

Un exemplu de descriere a conţinutului de culoare este prezentat în Figura 3.2 unde sunt ilustrate histogramele de culoare pentru imagini de sport şi respectiv animaţie (în fiecare caz sunt ilustrate câteva imagini reprezentative). Histograma de culoare este calculată folosind metoda propusă în [Ionescu 11] culorile fiind proiectate la paleta Webmaster de 216 culori. Se poate observa faptul că decriptorul de culoare astfel creat ilustrează particularităţile fiecărui tip de conţinut, imaginile de sport au o tentă predominant verde în timp ce imaginile de animaţie sunt predominant galbene-portocaliu conform conţinutului acestora.

[2]paleta de culoare a unei imagini reprezintă mulţimea tuturor culorilor prezente în această imagine. Aceasta reprezintă o sub-mulţime a spaţiului de culoare în care este reprezentată imaginea.

[3]`http://en.wikipedia.org/wiki/8-bit_color`

[4]`http://www.visibone.com/colorlab`

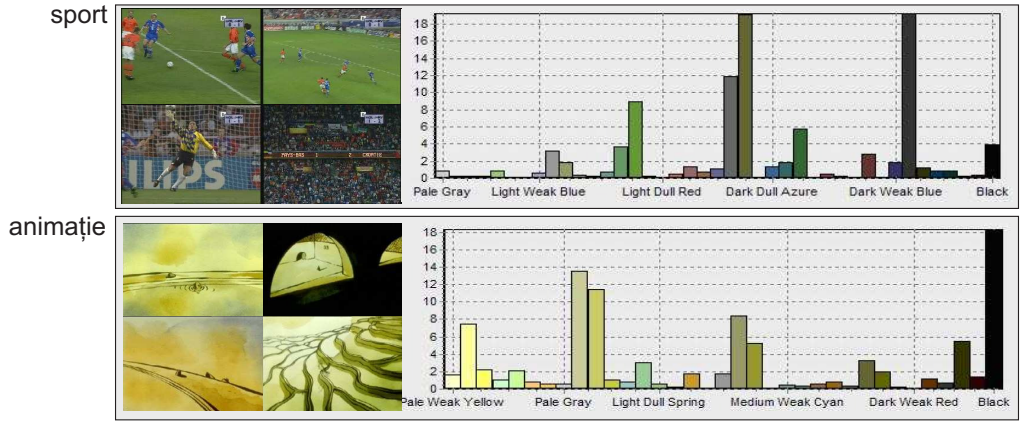


Figura 3.2: Exemplu de descriere a culorilor folosind histograme de culoare în cazul imaginilor de fotbal şi respectiv de animaţie (pe axa orizontală sunt reprezentate culorile în timp ce valorile de pe axa verticală sunt proporţionale cu procentul de apariţie al acestora în imagini) [Ionescu 11].

Informaţia relativă la forme se referă la caracterizarea proprietăţilor obiectelor prezente în scenă din perspectiva proprietăţilor geometrice ale acestora, fiind specifică imaginilor. Analiza formelor presupune detecţia în prealabil a obiectelor din scenă ce este realizată folosind tehnici de segmentare bazate pe contur sau pe regiuni de pixeli [Jain 89]. Succesul adnotării este astfel direct condiţionat de calitatea segmentării imaginii.

Problema descrierii formelor nu este una simplă în principal datorită faptului că imaginea nu este altceva decât o proiecţie bidimensională a lumii 3D, ceea ce înseamnă că una dintre dimensiunile obiectelor este pierdută. Astfel, formele extrase din imagine vor reprezenta numai parţial informaţia reală din scenă. Mai mult, imaginea este perturbată de zgomot[5] şi defecte de achiziţie ceea ce cresc dificultatea obţinerii unei reprezentări robuste.

Un descriptor de formă trebuie să fie eficient în sensul în care acesta trebuie să furnizeze suficientă putere discriminatorie pentru a identifica obiectele similare perceptual în contextul în care acestea pot fi reprezentate în contexte diferite (de exemplu scene diferite, momente temporale diferite), din diverse unghiuri, distorsionat, parţial sau suprapuse peste alte obiecte.

Acest lucru presupune atât o invarianţă la zgomot, cât şi la rotaţii, translaţii, modificări de scală sau în general la orice tip de transformare

[5]zgomotul în imagine se referă la acea informaţie perturbatoare ce alterează informaţia utilă.

afină[6]. Problema ocluziei obiectelor poate fi rezolvată prin integrarea de informaţii suplimentare, precum o evoluţie temporală a imaginilor sau informaţie de adâncime (informaţie 3D).

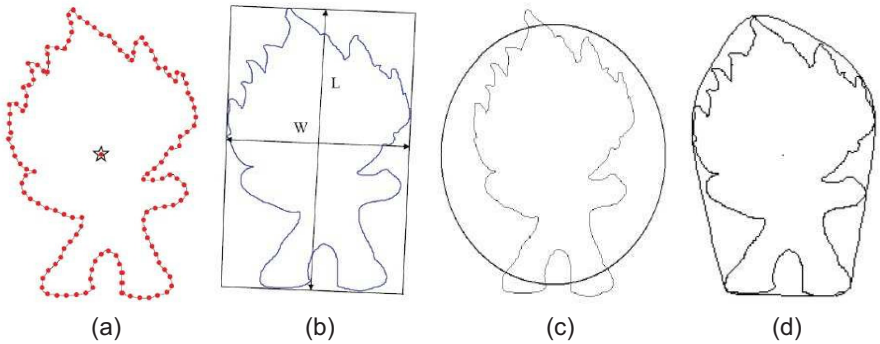


Figura 3.3: Exemplu de descriere a formelor: (a) reprezentarea centrului de greutate pe baza eşantionării uniforme a conturului, (b) determinarea parametrilor de elongaţie în funcţie de rata de aspect a formei (W/L), (c) determinarea raportului de circularitate (arie obiect raportat la aria cercului de acelaşi perimetru), (d) convexitate (cea mai mică regiune convexă ce include obiectul). Sursă imagini [Mingqiang 08].

Descriptorii de formă sunt calculaţi fie folosind doar informaţia de contur exterior a obiectelor sau informaţia de contur în relaţie cu informaţia din interiorul obiectului (regiunea plină a obiectului). Abordările existente variază de la calculul unor parametri simpli precum suprafaţă, orientarea axelor principale ale obiectului, convexitate, curbură, lungime, la parametri mai complexi precum momente statistice invariante, parametri spectrali (Fourier sau wavelet), reprezentarea sub formă de coduri (descompunerea conturului în secvenţe de segmente de dimensiune unitate şi codarea acestora), descompunerea în poligoane, reprezentări de tip "scale-space" (conturul este caracterizat la mai multe niveluri de scală), reprezentare cu matrice de forme, şi aşa mai departe [Mingqiang 08]. Un studiu detaliat este prezentat în [Smeulders 00]. O serie de exemple sunt prezentate în Figura 3.3.

Informaţia de textură. Conceptul de textură este legat de caracteriza-

[6]o transformare afină (cuvântul "affinis" în Latină înseamnă "conectat cu") reprezintă o transformare geometrică ce are proprietatea de a păstra coliniaritatea punctelor precum şi a rapoartelor de distanţă dintre punctele ce se găsesc pe o aceeaşi dreaptă (de exemplu, punctul de mijloc al unei drepte în urma transformării îşi va păstra proprietatea). O transformare afină nu garantează totuşi conservarea unghiurilor sau a lungimilor, dar are proprietatea de a păstra liniile paralele.

rea proprietăţilor materialelor prezente în imagini şi presupune atât analiza informaţiei de culoare cât şi de contur. O textură este definită ca fiind o regiune din imagine ce prezintă caracteristici omogene, precum un motiv de bază ce se repetă în domeniul spaţial sau frecvenţial.

Un exemplu este ilustrat în Figura 3.4. Tehnicile de descriere a texturilor presupun cuantificarea acestor proprietăţi pentru a caracteriza o serie de atribute specifice, precum asperitate, uniformitate, variabilitate, direcţionalitate, regularitate, ca o funcţie de variaţia spaţială a intensităţii pixelilor din imagine (de regulă exprimată ca niveluri de gri). Metodele existente pot fi clasificate în abordări statistice, geometrice, pe bază de modele şi pe bază de filtre [Tuceryan 93].

Figura 3.4: Exemplu de texturi (de la stânga la dreapta şi de sus în jos): perete de cărămidă, parchet lemn, ciment, pavaj piatră, zid de piatră, structură osoasă, pavaj de piatră radial şi textură artificială (sursă imagini Wikipedia).

Una dintre cele mai utilizate abordări o constituie metodele statistice. Distribuţia spaţială a intensităţii pixelilor este caracterizată statistic, ca de exemplu prin calcularea probabilităţii de co-ocurenţă a unei anumite intensităţi în diverse direcţii şi distanţe faţă de un punct de referinţă. Statisticile pot fi calculate pentru valorile unui singur pixel (statistici de ordinul întâi) sau pentru perechi sau regiuni de pixeli (statistici de ordin superior). Astfel de exemple sunt parametrii extraşi din matricele de co-ocurenţă (de exemplu: energie, contrast, corelaţie), parametrii de autocorelaţie sau histogramele de contur.

Abordările geometrice analizează textura din perspectiva proprietăţilor geometrice ale primitivelor acesteia (elementele texturii) precum arie, formă, lungime şi a modului de distribuţie al acestora într-o anumită reţea (sau "grid"). De exemplu, imaginea unui zid de cărămidă poate fi descrisă pe baza unei singure cărămizi (primitiva texturii în acest caz) şi prin definirea reţelei de plasare a acesteia în spaţiu.

O altă categorie de abordări sunt metodele bazate pe modele. Texturile sunt sintetizate pe baza unui model al cărui parametrii descriu proprietăţile esenţiale ale acesteia. De exemplu, elementele texturii pot fi modelate ca puncte întunecate sau luminoase, ca tranziţii verticale sau orizontale, ca linii. Exemple de astfel de modele sunt lanţurile Markov[7] şi modelarea fractală[8].

Metodele bazate pe filtre sunt specifice domeniului prelucrării de semnal. Acestea se bazează practic pe filtrarea imaginii atât în domeniul spaţial cât şi frecvenţial. Dintre filtrele cel mai des utilizate sunt operatorii de derivare (de exemplu Laplacian, Roberts) sau filtrele Gabor[9]. Un studiu detaliat al literaturii este prezentat în [Smeulders 00].

Informaţia de mişcare. Conceptul de mişcare este definit în contextul secvenţelor de imagini, numite şi imagini în mişcare. O secvenţă de imagini presupune o evoluţie temporală a conţinutului unei imagini (informaţie spaţio-temporală; în cazul în care se adaugă şi informaţie audio obţinem ceea ce numim video - informaţie audio-vizuală). Dacă considerăm standardul de codare video PAL - Phase Alternating Line (unul dintre cele mai răspândite în Europa) o secundă dintr-o secvenţă video corespunde la o succesiune de nu mai puţin de 25 de imagini. Caracterizarea informaţiei de mişcare presupune astfel caracterizarea schimbărilor (de regulă spaţiale) ce au loc de la o imagine la alta. Aceste schimbări pot fi analizate local, doar pentru o anumită regiune din imagine (de exemplu mişcarea unui obiect în scenă), sau global pentru întreaga imagine (de exemplu mişcarea camerei video).

Pentru a putea descriere conţinutul de mişcare este nevoie mai întâi de realizarea unei etape intermediare ce presupune identificarea acestuia în secvenţă. O abordare simplificată presupune detecţia mişcării [Bovik 09]. Aceasta are ce are ca scop localizarea acelor regiuni de pixeli din imagine în care survin schimbări în timp, de regulă de la o imagine la alta. Limitarea acestei abordări constă în faptul că nu se ţine cont de natura acestor schimbări, acestea putând surveni, în special în cazul secvenţelor editate în studio, independent de mişcare, de exemplu prin fluctuaţii de intensitate,

[7]un lanţ Markov (denumit după Andrey Markov) reprezintă un sistem matematic caracterizat de tranziţii succesive între un număr finit, măsurabil, de stări posibile. Acesta este un proces aleator fără memorie în sensul în care tranziţia sistemului la o altă stare depinde doar de starea curentă şi nu depinde de stările anterioare.

[8]un fractal (termen creat de Benoît Mandelbrot, din Latină "fractus" - neregulat) reprezintă o suprafaţă de formă neregulată sau fragmentată creată pe baza unor reguli deterministe sau stohastice ce implică un proces de omotetie internă (transformare geometrică în care punctele corespondente sunt coliniare cu un punct fix (centru), distanţa faţă de el crescând sau reducându-se în raport constant - sursă "Marele Dicţionar de Neologisme").

[9]un filtru Gabor (denumit astfel după Dennis Gabor) reprezintă un filtru liniar ce are proprietatea de a avea caracteristici similare filtrelor din sistemului vizual uman.

efecte speciale.

Exemple de astfel de metode includ detecţia cu prag fix sau adaptiv (o regiune este declarată de mişcare dacă diferenţele dintre pixeli pentru două imagini succesive sunt mai mari decât un anumit prag), tehnici de estimare a fundalului[10] precum media alunecătoare, aproximare filtru median, metode statistice neparametrice, metode recursive şi aşa mai departe. Un exemplu de detecţie este ilustrat în Figura 3.5.(b).

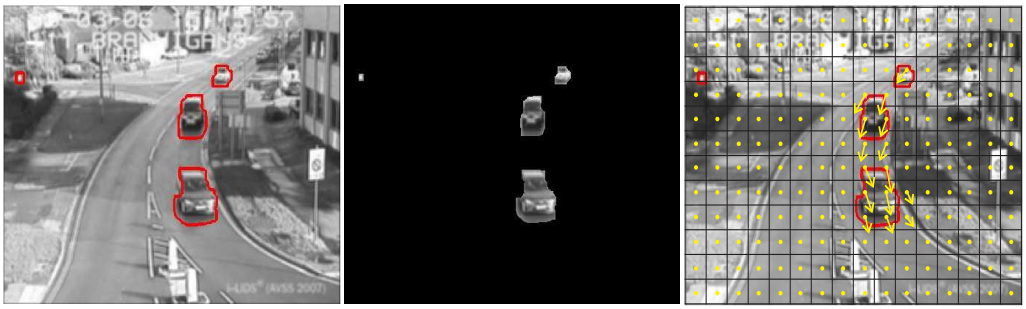


Figura 3.5: Exemple de determinare a conţinutului de mişcare pentru o secvenţă de supraveghere video: (a) imagine din secvenţa originală (obiectele care se deplasează sunt încercuite cu roşu), (b) detecţie de mişcare folosind aproximarea filtrului median (imaginea reprezintă regiunile care se schimbă), (c) câmpul vectorial de mişcare obţinut cu o estimare pe blocuri de pixeli (vectorii de mişcare sunt ilustraţi cu galben, punctele semnifică absenţa mişcării).

O a doua abordare o constituie tehnicile de estimare a mişcării [Bovik 09]. Acestea, spre deosebire de detecţia mişcării, presupun estimarea deplasărilor pixelilor sau a regiunilor de pixeli de la o imagine la alta, estimare ce este cuantificată prin asocierea unui vector de mişcare. Acesta indică atât direcţia deplasării pixelilor (orientare) cât şi deplasarea spaţială (amplitudine). În urma estimări, imaginea este practic reprezentată de un câmp de astfel de vectori de mişcare indicând modul de deplasare al fiecărui pixel sau bloc de pixeli. Un exemplu este prezentat în Figura 3.5.(c).

Tehnicile de estimare a mişcării variază de la abordări bazate pe metode diferenţiale (bazate pe estimarea fluxului optic), parametrice, stohastice sau bazate pe blocuri ("block-based") acestea din urmă regăsindu-se în toate standardele de codare video precum cele dezvoltate de Moving Picture Ex-

[10]detecţia fundalului sau "background" presupune localizarea acelor pixeli din imagine ce rămân aproximativ constanţi de la o imagine la alta.

perts Group - MPEG[11] (informaţia relativă la deplasarea regiunilor permite reconstrucţia imaginilor, ceea ce oferă un factor de compresie semnificativ).

Odată identificat conţinutul de mişcare, descriptorii de conţinut cuantifică o serie de proprietăţi ale acestuia. Ca exemple de descriptori putem enumera determinarea traiectoriei obiectelor din scenă, identificarea tipului de mişcare a camerei video ("zoom" - apropiere/depărtare, rotaţie, translaţie), determinarea activităţii de mişcare folosind cuantizarea varianţei amplitudinii vectorilor de mişcare, determinarea distribuţiei spaţiale şi temporale a activităţii de mişcare, construcţia de imagini MHI de "istorie a mişcării" (Motion History Images) formate prin acumularea informaţiei de mişcare a fiecărui pixel într-o anumită fereastră temporală, determinarea de histograme de intensitate şi aşa mai departe. De menţionat faptul că determinarea descriptorilor de mişcare depinde de succesul şi calitatea detecţiei/estimării de mişcare folosită.

Informaţia de structură temporală. Decriptorii de conţinut relativi la structura temporală se adresează secvenţelor de imagini şi în special secvenţelor editate în studio, precum filme, reportaje, sport şi aşa mai departe (în general materiale destinate distribuţiei TV).

Descrierea structurii temporale video implică segmentarea temporală a acesteia prin descompunerea secvenţei în unităţi structurale de bază numite şi plane video [Lienhart 01]. Un plan video este practic o secvenţă de imagini întregistrată între pornirea şi oprirea camerei video având proprietăţile de unitate temporală şi de loc (vezi Figura 3.6). Pentru a obţine secvenţa finală, planele video sunt concatenate folosind diverse tranziţii video. O tranziţie video nu este altceva decât un efect vizual ce poate presupune fie o tranziţie abruptă de tip "cut" (concatenarea directă a două plane succesive), fie tranziţii graduale precum "fades" (apariţia sau dispariţia imaginii dintr-o imagine constantă, de regulă neagră), "dissolves" (transformarea graduală a unei imaginii în alta), "mattes", "wipes" şi aşa mai departe [Bimbo 99]. Câteva exemple sunt ilustrate în Figura 3.6.

Practic segmentarea temporală implică localizarea în secvenţă a acestor tranziţii. Ca frecvenţă de apariţie, tranziţiile de tip "cut" sunt cele mai frecvente, de regulă 30 de minute video pot conţine până la 300 de astfel de tranziţii în timp ce frecvenţa tranziţiilor graduale este cu cel puţin un ordin

[11]MPEG sau Moving Picture Experts Group, reprezintă o organizaţie internaţională ce se ocupă cu dezvoltarea normelor pentru compresia, decompresia şi analiza şi codarea video. Aceasta este responsabilă pentru dezvoltarea standardelor clasice de codare, precum MPEG-1 folosit pentru formatul VideoCD, MPEG-2 folosit pentru stocarea pe DVD, MPEG-4 folosit la BD (Blu-Ray Disc), MPEG-7 standard de descriere a conţinutului video pentru indexare multimedia sau MPEG-21 ce defineşte interoperabilitatea tuturor tipurilor de conţinut multimedia.

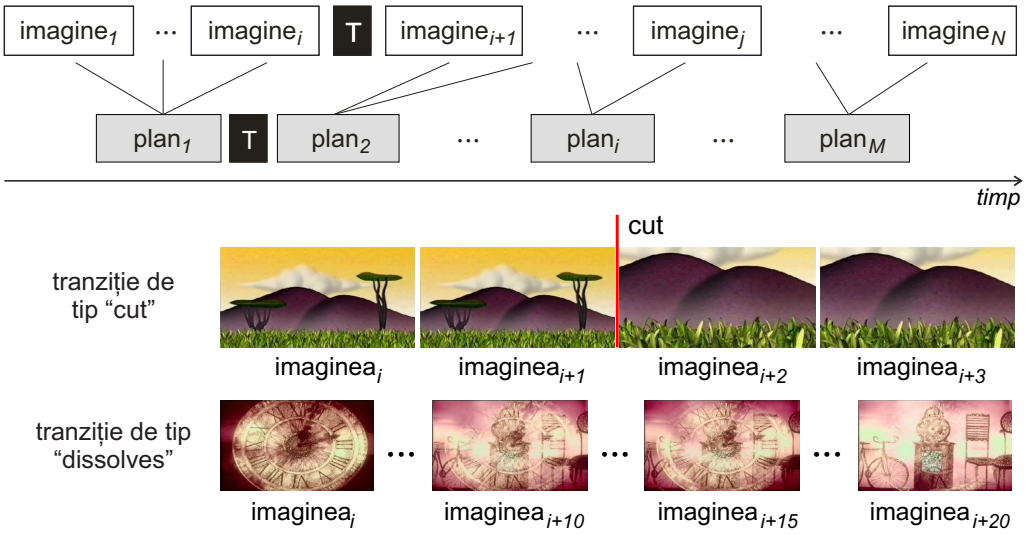


Figura 3.6: Structura temporală a unei secvenţe video (T reprezintă o tranziţie video, N este numărul de imagini al secvenţei iar M numărul de plane video). În partea de jos a imaginii sunt ilustrate un exemplu de tranziţie de tip "cut" (imagini film de animaţie "Gazoon" [CITIA 13]) şi respectiv "dissolve" (imagini film de animaţie "Coeur de Secours" [CITIA 13]).

de mărime mai redusă.

Metodele de detecţie a tranziţiilor video exploatează în general detecţia discontinuităţii vizuale produse de acestea în fluxul video folosind abordări bazate pe analiza intensităţii pixelilor (de exemplu o tranziţie de tip "cut" implică o diferenţă semnificativă a distribuţiei de culoare ce poate fi analizată folosind diferenţa dintre histograme, o tranziţie de tip "fade" implică o variaţie graduală a intensităţii luminoase), analiza contururilor (de exemplu o tranziţie de tip "dissolves" presupune un raport semnificativ de puncte de contur ce apar/dispar din imagine), analiza mişcării (de exemplu o tranziţie de tip "cut" produce o discontinuitate a vectorilor de mişcare) sau analiza informaţiei în domeniul comprimat (precum analiza coeficienţilor transformatei cosinus discrete din fluxul MPEG).

La un nivel de descriere superioară, segmentarea secvenţei poate implica descompunerea acesteia în unităţi structurale de nivel semantic superior, precum gruparea planelor video în scene (grupuri de plane video ce sunt corelate din punct de vedere al conţinutului semantic şi presupun unitate de loc, de timp şi de acţiune), în episoade (grupuri de scene ce sunt similare din punct de vedere al acţiunii globale, ca de exemplu episoadele unei serii TV) şi aşa mai departe.

Interesul în segmentarea temporală este dublu. Pe de-o parte, acesta constituie primul pas de analiză pentru marea parte a metodelor de analiză a conţinutului video deoarece furnizează informaţii relative la structura semantică a acestuia. De exemplu, având la dispoziţie structura de plane sau de scene video, analiza de conţinut se poate realiza în interiorul acestora evitând astfel prelucrarea imaginilor de tranziţie cât şi asigurând unitatea semantică a conţinutului.

Extragerea unui descriptor pentru un segment ales aleator din secvenţă riscă să amestece informaţii distincte. De exemplu, dacă considerăm cazul particular al unei secvenţe de ştiri şi segmentul ales conţine atât înregistrarea prezentatorului cât şi a unui reportaj extern, amestecarea informaţiilor vizuale ale celor două subiecte complet diferite nu poate produce un descriptor reprezentativ.

Pe de altă parte, structura temporală furnizează ea însăşi informaţii de conţinut. Folosirea unui anumit tip de tranziţii pentru a face legătura între planele video nu este aleatorie ci corespunde unor reguli cinematice de montaj bine definite [Reynertson 70]. De exemplu, folosirea frecventă a tranziţiilor de tip "cut" are ca efect creşterea dinamismului secvenţei, tranziţiile de tip "dissolve" şi "fade" sunt folosite frecvent pentru a schimba timpul sau locul acţiunii, o secvenţă de tip "fade-out - fade-in" introduce un moment de pauză în derularea acţiunii ca de exemplu pentru a trece la un alt capitol al naraţiunii.

Descriptorii ce caracterizează informaţia de structură temporală exploatează în principal frecvenţa de apariţie a schimbărilor de plan video în secvenţă, fie în mod direct prin măsuri precum determinarea duratei medii a planelor video, ratei medii de schimbare de plan raportată la unitatea temporală (de regulă denumită ritm vizual), raportului tranziţiilor graduale din secvenţă, extragerea de imagini cheie la nivelul fiecărui plan şi prelucrarea acestora folosind descriptori clasici de imagine (vezi secţiunile anterioare); fie derivând informaţii relative la activitatea vizuală a secvenţei exploatând conceptul de acţiune (concept de regulă asociat frecvenţei de tranziţii de tip "cut", de exemplu o secvenţă de acţiune va avea o densitate ridicată de plane video de scurtă durată în timp ce o secvenţă a unui documentar este foarte probabil să conţină doar câteva plane video).

Trăsături. Informaţia legată de ceea ce numim trăsături ("features") este de fapt un caz particular de descriere a informaţiei de contur în imagini şi este strâns legată de noţiunea de puncte de interes ("interest points").

Un punct de inters în imagine reprezintă de regulă o regiune de pixeli (de dimensiuni reduse) a căror proprietăţi o fac reprezentativă pentru întelegerea conţinutului structural al imaginii. Nu orice regiune de pixeli care conţine

contururi este astfel un punct de interes.

De exemplu, dacă considerăm sistemul vizual uman, se ştie faptul că ochiul este mai sensibil la percepţia punctelor de inflexiune din imagini, precum unghiuri sau intersecţii, decât la informaţiile redundante, continuue, precum liniile drepte. Aceste informaţii sunt acelea ce tind să fie percepute primele în imagine, fiind definitorii, şi apoi pe baza lor să se realizeze o aproximare a scenei.

În Figura 3.7.(a) am ilustrat un exemplu în acest sens. În imagine sunt reprezentate patru treimi de cercuri dispuse simetric. La o primă vedere ochiul uman tinde să perceapă mai întâi cele patru colţuri contrastate (puncte de interes) şi să extrapoleze informaţia imaginând un dreptunghi alb suprapus peste patru cercuri negre. Totuşi în realitate, liniile ce definesc dreptunghiul nu există, fiind doar o iluzie.

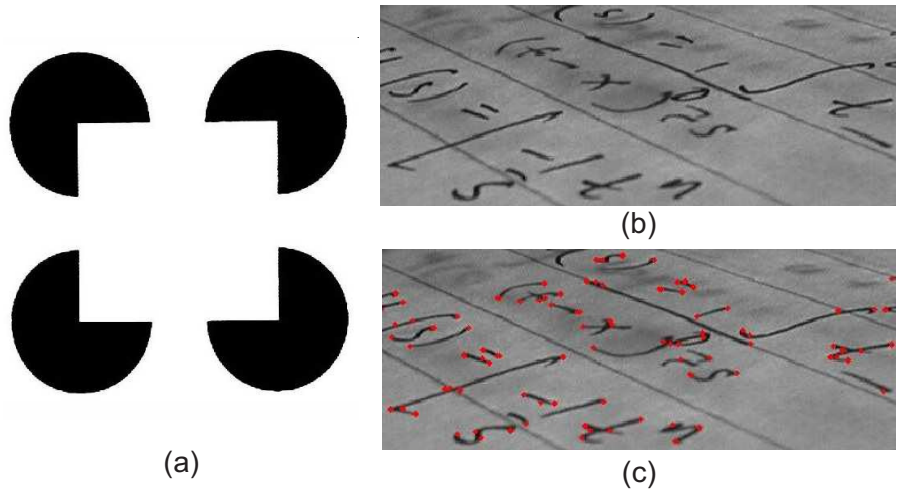


Figura 3.7: Exemplu de trăsături: (a) exemplu de iluzie optică în care cele patru treimi de cerc sunt percepute ca un dreptunghi suprapus peste patru cercuri negre (sursă http://webvision.med.utah.edu/book/), (b) şi (c) ilustrează un exemplu de detector de colţuri, prima imagine fiind imaginea iniţială iar în a doua imagine punctele roşii marchează trăsăturile detectate (sursă imagini Wikipedia).

În contextul imaginilor, aceste trasături pot fi formalizate ca fiind acele puncte din imagine ce întrunesc următoarele proprietăţi: au o definiţie matematică bine precizată, au o poziţie bine definită în imagine, informaţia locală din jurul punctului de interes este bogată informaţional (cu alte cuvinte sunt definite de context), şi cea mai importantă proprietate este aceea că acestea

trebuie să fie stabile la perturbaţii locale şi globale precum deformări datorate transformărilor de perspectivă, schimbarea unghiului de vizualizare, schimbări de scală, rotaţii, translaţii cât şi variaţii de iluminare (de exemplu: colţul unui dreptughi işi va păstra proprietatea indiferent dacă este întunecat, rotit, micşorat sau schimbată perspectiva). Datorită acestor proprietăţi, punctele de interes sunt de departe cea mai eficientă modalitate de reprezentare a conţinului imaginilor în contextul indexării după conţinut.

Tehnicile de detecţie a punctelor de interes/trăsături au pornit iniţial de la ideea detectării colţurilor în imagini, un astfel de detector fiind "Harris corner detector" ce foloseşte ipoteza conform căreia gradienţii (diferenţele) pe cele două direcţii oX şi respectiv oY trebuie să fie ambele semnificative pentru un colţ. Un exemplu de astfel de detecţie este prezentat în Figura 3.7.(c).

Alte metode mai elaborate sunt detectorul Harris Laplace (cunoscut ca detectorul Harris multi-scală) ce adaugă localizarea colţurilor folosind reprezentării ale imagini pe diverse niveluri de scală, detectorul Hessian Laplace, abordări ce folosesc reprezentări de tip "scale-space" (similar informaţiei de contur) precum Laplacian of Gaussian (LoG), Difference of Gaussian (DoG) sau Determinant of Hessian (DoH), detectorul Maximally Stable Extremum Regions (MSER) ce selectează anumite regiuni conexe din imagine dacă acestea sunt stabile în urma filtării repetate cu diverse praguri ("thresholding"[12]), până la binecunoscuţii detectori Scale Invariant Feature Transform - SIFT (ce se bazează pe localizarea maximelor şi minimelor obţinute în urma aplicării unor funcţii de diferenţe de Gaussiene[13]) şi respectiv Speeded Up Robust Features - SURF (ce foloseşte o descompunere wavelet de tip Haar şi imagini integrale). O trecere în revistă a diferitelor tehnici de analiză a trăsăturilor în imagini poate fi consultată în [Gauglitz 11].

Având în vedere eficienţa descriptorilor de trăsături în reprezentarea structurii imagini, în special datorată învarianţei acestora la o gamă largă de transformări, cercetările actuale în domeniu vizează extensia acestora pentru a putea exploata conţinutul temporal specific secvenţelor de imagini.

Dintre descriptorii de trăsături spaţio-temporali putem menţiona "Har-

[12]"thresholding" în prelucrarea de imagini reprezintă operaţia prin care valorile imaginii sunt transformate prin compararea cu un simplu prag de regulă obţinând o imagine binară. Dacă valoarea din imagine este superioară pragului aceasta este schimbată intr-o constantă (de regulă 1) şi în caz contrar într-o altă constantă (de regulă 0).

[13]diferenţa de Gaussiene constă în realizarea diferenţei dintre două variante înceţoşate ale imaginii iniţiale, de regulă prima imagine fiind mai înceţoşată ("blurred"). Înceţoşarea unei imagini presupune înlăturarea frecvenţelor înalte (de exemplu zone ne-uniforme precum contururi). Prin realizarea diferenţei între două astfel de imagini se obţine un filtru trece bandă care conservă doar o gamă de frecvenţe spaţiale din imaginea initială şi astfel doar anumite informaţii din imagine.

ris3D corner detector" (extensie a detectorului de colţuri pentru a include pe lângă gradienţii spaţiali şi gradienţi temporali), detectorul Cuboid ce foloseşte filtre Gabor temporale (vezi explicaţia de la informaţia de textură) pentru a detecta acele trăsături cu proprietăţi spaţiale particulare şi ce presupun o mişcare complexă, detectorul Hessian 3D ce se bazează pe estimarea determinantului matricei Hessiene[14] în care derivatele parţiale sunt calculate şi temporal, tehnici de eşantionare densă precum extragerea de regiuni 3D din secvenţă (de exemplu o porţiune de imagine pentru mai multe momente de timp) şi descrierea acestora adaptând descriptori de trăsături precum SURF 3D (extensie a descriptorului SURF). Un studiu detaliat al descriptorilor spaţio-temporali este prezentat în [Stöttinger 10].

3.2 Informaţia audio

Informaţia audio reprezintă o altă sursă importantă de informaţii relative la conţinutul datelor multimedia. Aceasta se referă la caracterizarea sunetului, fie în contextul video unde acesta este sincronizat informaţiei vizuale, fie independent (de exemplu fişiere audio de muzică, înregistrări, etc.). În general sunt vizate analiza şi identificarea vorbirii, a zgomotului şi a efectelor sonore sau analiza conţinutului muzical.

Prelucrarea semnalului audio se realizează principial într-un mod similar prelucrării secvenţelor de imagini fiind de asemenea o reprezentare temporală a datelor. Un semnal audio digital (discret) nu este altceva decât o secvenţă de eşantioane (valori de amplitudine ale undelor sonore) înregistrate în timp (vezi Figura 3.8.(a)). Acestea sunt prelucrate la nivel de cadre audio, un cadru audio fiind o secvenţă temporală ce conţine un anumit număr de eşantioane (un exemplu de valoare uzuală este folosirea a 1024 de eşantioane). Important este faptul că aceste cadre nu sunt întotdeauna disjuncte, de regulă fiind suprapuse cu până la 50% din durată. Acest lucru asigură faptul că toate părţile semnalului audio vor fi bine reprezentate la nivel de cadre.

Metodele de descriere a conţinutului audio se împart în două categorii. Metode ce analizează informaţia audio direct în domeniul temporal la nivel de cadru sau folosind o reprezentare statistică a distribuţiei acestora în documentul audio (descriptorii extraşi la nivel de cadru sunt agregaţi pentru întreaga secvenţă prin statistici de medie, varianţă, median şi aşa mai departe). Dintre descriptorii cei mai frecvent folosiţi în acest caz putem

[14]matricea Hessiană (după numele matematicianului Ludwig Otto Hesse) reprezintă matricea pătratică a derivatelor parţiale de ordin doi ale unei anumite funcţii de mai multe variabile. Definită în acest fel, aceasta are proprietatea de a descrie curbura locală a funcţiei.

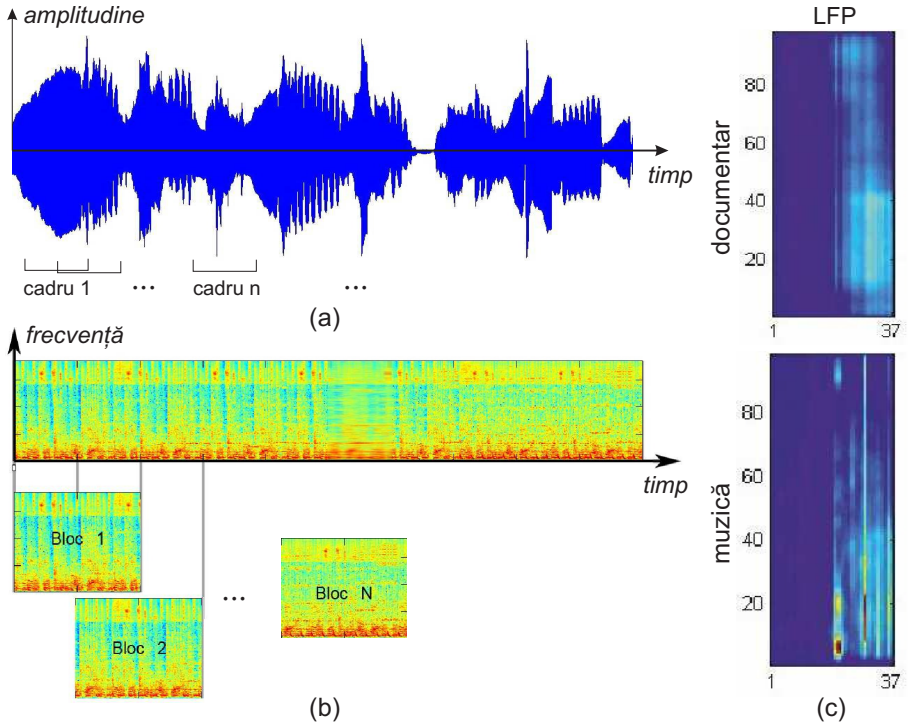


Figura 3.8: Exemplu de prelucrare a semnalului audio: (a) analiză în domeniul temporal, (b) analiză pe blocuri de cadre spectrale în domeniul frecvenţial [Seyerlehner 10], (c) exemplu de descriptor Logarithmic Fluctuation Pattern (LFP) [Ionescu 12b] în cazul unui documentar şi videoclip muzical (la acesta din urmă se observă aspectul ritmic prezent prin maximele LFP - reprezentate cu roşu şi galben).

menţiona: Zero-Crossing Rate (ZCR) ce reprezintă numărul de treceri prin zero ale semnalului raportat pe unitatea de timp, energia semnalului (Root Mean Square of Signal Energy sau RMS), rata de absenţă a sunetului sau coeficienţii de autocorelaţie ai semnalului [Mathieu 10].

Totuşi marea parte a metodelor analizează sunetul în domeniul frecvenţial. Pentru aceasta, fiecare cadru audio este reprezentat în domeniul transformatei Fourier iar informaţia obţinută este prelucrată într-o reprezentare frecvenţă (dată de reprezentarea Fourier a cadrelor audio) - timp (dată de succesiunea cadrelor audio în timp) ca de exemplu folosind spectograma de aplitudine - reprezentarea temporală a amplitudinii transformatei Fourier a fiecărui cadru audio.

Dintre abordările folosite putem menţiona distribuţia energiei semnalului, centroizii frecvenţelor, lărgimea de bandă, "pitch", "loudness" sau reprezen-

tarea coeficienţilor cepstrali Mel-Frequency Cepstral Coefficients (MFCC). La rândul ei, reprezentarea semnalului frecvenţă-timp poate fi prelucrată pe blocuri de cadre spectrale (valori uzuale sunt de ordinul a 10 pana la 512 cadre per bloc) ceea ce are avantajul de a integra pe lângă informaţia de frecvenţă şi informaţie temporală locală, ca de exemplu aspecte ritmice ale semnalului (vezi Figura 3.8.(b)).

Dintre descriptorii de acest gen putem menţiona Spectral Pattern (informaţie relativă la timbru sonor), Logarithmic Fluctuation Pattern (informaţie relativă la aspectele ritmice ale semnalului), Correlation Pattern (informaţie relativă la schimbările de intensitate) sau Spectral Contrast Pattern (informaţie relativă la tonalitate) [Seyerlehner 10].

Un exemplu este prezentat în Figura 3.8.(c) în care am ilustrat descriptorul Logarithmic Fluctuation Pattern (LFP) în cazul coloanei sonore a unui documentar şi respectiv a unui videoclip muzical. O caracteristică specifică muzicii este prezenţa de bătăi ritmice ("beats") ce sunt vizibile sub formă de maxime ale LFP (vezi zone colorate cu roşu şi galben). În contrast, în cazul documentarului, structura LFP este plată ceea ce indică că nu există elemente repetitive percutante în fluxul audio.

O altă direcţie de studiu importantă ce vizează analiza sunetului o reprezintă tehnicile de recunoaştere automată a vorbirii (Automatic Speech Recognition sau ASR [Lamel 08]). Acestea au ca obiectiv transformarea vorbirii din semnal audio în text ce poate fi prelucrat mai departe folosind tehnici specifice. Folosirea textului obţinut în urma ASR furnizează informaţii preţioase relative la conţinutul datelor. Avantajul descriptorilor textuali precum şi limitările ASR sunt discutate în secţiunea următoare.

Raportat la descriptorii vizuali, descriptorii audio tind să furnizeze o putere discriminatorie mai bună în marea parte a aplicaţiilor relative indexării după conţinut, precum identificarea genului video sau detecţia anumitor concepte video [Ionescu 12b] [Over 12].

3.3 Informaţia textuală

De departe cea mai eficientă sursă de informaţii pentru indexare o constituie textul. Aproape în totalitate, sistemele existente de căutare multimedia se bazează pe descriptori textuali. Avantajul reprezentării textuale este acela că oferă un nivel de descriere semantică a conţinutului, foarte apropiat de nivelul de percepţie uman. Mai mult, exprimarea textuală este la îndemâna oricărui utilizator, ceea ce rezolvă problematica formulării cererilor de căutare.

Totuşi dezavantajul principal al informaţiei textuale este dat de posiblitatea limitată de automatizare a procesului de generare, aceasta necesitând

practic să fie furnizată de utilizator. De exemplu, la încărcarea on-line a unei imagini, de regulă utilizatorul va specifica o scurtă descriere textuală a conţinutului acesteia, ca de exemplu "Turnul din Pisa". Aceasta va fi folosită ulterior drept descriptor de conţinut pentru indexare. Totuşi această informaţie este incompletă şi nu descrie decât global conţinutul, nu există informaţii relative la persoanele din scenă, la momentul zilei sau relativ la prezenţa altor obiecte. Acest lucru limitează această imagine să nu poată fi găsită decât în cazul căutării unor imagini cu turnul din Pisa, şi nu pentru alte informaţii din scenă.

În cazul datelor multimedia, informaţia textuală poate fi obţinută din mai multe surse. Conform celor menţionate anterior, o primă sursă de descriptori textuali este însuşi utilizatorul, aceste date fiind generate manual. În acest caz, informaţia textuală este de regulă reprezentată sub formă de mici rezumate de conţinut referitoare la date ("synopsis", de exemplu în cazul unui film acestea pot fi rezumatul naraţiunii acestuia), etichete de conţinut ("user tags" ce reprezintă de regulă câteva cuvinte cheie ce descriu conţinutul global), subtitrări în cazul filmelor, metadate[15] (ce furnizează o serie de informaţii suplimentare de tipuri diferite, legături ("link") către alte surse de informaţii, proprietăţi ale datelor), informaţii referitoare la localizarea geografică a datelor precum coordonatele GPS[16] ale unei imagini (longitudine, latitudine), comentarile utilizatorilor relativ la conţinutul datelor specifice de regulă reţelelor de socializare sau textul ce înconjoară elementul multimedia respectiv pe o pagină web.

Câteva exemple de astfel de descrieri au fost prezentate în Figura 3.1 în care am ilustrat o pagină tipică de pe platforma YouTube. Se pot observa diferitele informatii textuale, de la descrieri, tag-uri până la comentariile utilizatorilor asociate unei secvenţe video. Toate aceste informaţii sunt informaţii relative la conţinut.

O altă sursă de informaţie textuală o constituie textul conţinut chiar de datele multimedia. Avantajul acestuia il constituie faptul că poate fi extras folosind metode automate. O primă sursă este informaţia vizuală, ca de exemplu textul încrustat în imagine, scrisul de mână, subtitrările filmelor (în

[15]metadatele sunt definite uzual ca fiind "date despre date", sau altfel spus, date care descriu alte date, de orice fel şi de orice tip. Cu alte cuvinte, metadatele oferă informaţii suplimentare la o serie de date. De exemplu, o pagină web, pe lângă textul propriu-zis poate conţine metadate ce specifică limba în care este scrisă, modul de creare al paginii, diferite surse adiţionale de informaţii şi aşa mai departe.

[16]sistemul GPS - Global Positioning System reprezintă un sistem de poziţionare geografică bazat pe sateliţi ce furnizează informaţie de localizare şi timp independent de vreme şi pentru oricare poziţie de pe glob, atâta timp cât există posibilitatea de captare a semnalului de la cel puţin 3 sateliţi GPS.

cazul în care nu sunt disponibile separat), textul grafic (de exemplu diverse indicatoare precum denumirea unei străzi, numele unui obiect, numărul de înmatriculare al unei maşini, scorului în secvenţele sportive). Extragerea acestuia presupune folosirea de tehnici de recunoaştere automată a caracterelor sau OCR ("Optical Character Recognition"[17]).

O a doua sursă de text o constituie informaţia audio şi în special vorbirea (de exemplu naraţiune, dialoguri, monologuri). Aceasta poate fi recunoscută şi convertită în text folosind tehnicile de recunoaştere automată a vorbirii sau Automatic Speech Recognition (ASR) [Lamel 08]. Textul obţinut în acest fel oferă informaţii preţioase de conţinut, totuşi tehnicile de ASR sunt limitate pe de-o parte de diversitatea limbilor existente cât şi de imposibilitatea de a furniza o transcriere eficientă în condiţii de zgomot de fundal (cum este cazul filmelor).

Odată obţinută informaţia textuală aceasta poate fi folosită direct ca descriptor de conţinut. Totuşi în marea parte a cazurilor informaţia textuală tinde să fie redundantă şi de dimensiune semnificativă (de exemplu sute de mii de cuvinte) necesitând o reprezentare mai eficientă. Dintre metodele cel mai frecvent folosite putem enumera reprezentarea de tip Term Frequency–Inverse Document Frequency (TF–IDF) [Knees 09] şi Bag-of-Words (B-o-W) [Wallach 06].

TF–IDF este un model statistic ce se bazează pe determinarea gradului de importanţă al unui termen pentru un anumit document dintr-un corpus de date. Valoarea TF-IDF va creşte proporţional cu numărul de apariţii al termenului în document (term frequency) dar în acelaşi timp este compensată de frecvenţa de apariţie a cuvântului în corpus (inverse document frequency) ceea ce ajută la verificarea a cât de comun sau rar este termenul pentru toate documentele din corpus. Informaţia textuală poate fi astfel sintetizată prin valorile TF-IDF pentru un set de termeni cheie prestabiliţi în funcţie de aplicaţie sau extraşi chiar din document.

Modelul B-o-W este un model similar ce ţine cont de frecvenţa de apariţie a cuvintelor. În acest model textul este reprezentat sub forma unei colecţii, ne-ordonate, de cuvinte ("bag of words"), ignorând astfel orice reguli gramaticale. Pe baza acestei reprezentări se alcătuieşte mai întâi un dicţionar de cuvinte eliminând cuvintele care se repetă. Descriptorul texual pentru un anumit document va consta astfel în reprezentarea sub formă de histogramă a numărului de apariţii ale fiecărui cuvânt din dicţionar în documentul respectiv. Documentele ce descriu date similare vor avea o frecvenţă comparabilă

[17] recunoaşterea automată a caracterelor reprezintă procesul mecanic sau electronic de traducere a imaginilor ce conţin scris de mână, scris de maşină sau text imprimat (de regulă rezultate în urma procesului de scanare) în text editabil de către calculator.

a anumitor termeni.

3.4 Descriere semantică sau sintactică?

În general descriptorii de conţinut obţinuţi în urma adnotării conţinutului datelor multimedia pot fi clasificaţi în funcţie de nivelul semantic al informaţiilor furnizate în trei categorii:

Descriptori sintactici ("low-level"), constau de regulă în adnotarea datelor cu descrieri numerice. Acest mod de descriere corespunde în general primelor sisteme de indexare (cu toate acestea multe dintre metode sunt folosite şi în sistemele existente - vezi secţiunile anterioare). Adnotarea sintactică este definită generic ca fiind adnotarea ce se referă la *relaţiile dintre unităţile de nivel scăzut constituente ale datelor multimedia şi modul de constituire a structurii acestora*. Aceasta se poate realiza pe baza atributelor numerice, de nivel semantic redus, ca de exemplu parametri statistici calculaţi la nivel de pixel sau regiuni de pixeli, proprietăţi geometrice ale obiectelor, structura temporală a unei secvenţe sau vectori de mişcare. De regulă, descriptorii obţinuţi în urma procesului de adnotare sunt valori numerice ce descriu atribute de tipul celor enumerate mai sus dar şi relaţiile sintactice ce pot exista între acestea. Extraşi la acest nivel de percepţie, descriptorii sintactici sunt dificil accesibili utilizatorului de rând. De exemplu, căutarea unei imagini în funcţie de procentul de apariţie al unei culori sau a unei secvenţe de imagini care să conţină 30% mişcare de translaţie şi 20% mişcare de rotaţie, nu constituie o descriere prea relevantă pentru utilizator.

Descriptori simbolici ("mid-level"), aceştia corespund unui nivel de descriere intermediar, ce se găseşte între cele două extreme: numeric şi semantic, ca de exemplu denumirea culorilor dintr-o imagine, detectarea unei scene de dialog sau a prezenţei umane în scenă, identificarea unui anumit tip de conţinut. De regulă descriptorii de nivel semantic intermediar sunt determinaţi, indirect, pe baza descrierilor sintactice.

Descriptori semantici ("high-level"), în contrast cu adnotarea sintactică, adnotarea semantică a conţinutului presupune o descriere perceptuală ce tinde să atingă un nivel similar cu nivelul de percepţie uman. Informaţiile numerice obţinute în urma analizei sintactice pot fi convertite în concepte semantice precum conceptele lingvistice folosind informaţii "a priori" despre conţinutul datelor, şi/sau trecând printr-o etapă intermediară de descriere simbolică. Un sistem semantic este definit generic ca fiind *orice sistem ce implică o colecţie de simboluri (vocabularul sistemului), reguli ce permit con-*

stituirea de propoziţii, reguli de desemnare şi reguli de validare. În cazul sistemelor de indexare, termenul de "semantic" îşi conservă acest sens. Acesta se traduce prin *codarea interpretării datelor pentru a servi unei aplicaţii specifice* [Smeulders 00]. Astfel, descrierea semantică implică existenţa unui *set de simboluri* şi *reguli* ce permit interpretarea lingvistică a anumitor evenimente sau proprietăţi ale datelor multimedia.

Acest mod de descriere presupune dezvoltarea de tehnici capabile să furnizeze o înţelegere completă a conţinutului necesitând de cele mai multe ori o abordare multimodală (imagine-sunet-text). De exemplu, dacă ne limităm în a folosi doar informaţia furnizată de o imagine, să luăm cazul unei imagini ce surprinde un jucător de fotbal, singurele caracteristici ce reies din analiza imaginii sunt fizionomia acestuia şi prezenţa sa în scenă. Pe de altă parte, dacă dispunem de secvenţa ce îl surprinde pe jucător, putem determina dacă acesta va marca golul, modul în care acesta joacă, contextul înregistrării, cum ar fi meciul despre care este vorba şi aşa mai departe, informaţii semantice esenţiale pentru înţelegerea conţinutului secvenţei. În ciuda dificultăţii sporite de generare automată, acest mod de reprezentare al datelor este unul dintre cele mai eficiente şi constituie direcţia actuală de cercetare în domeniu.

Pentru a înţelege mai bine diferenţa dintre cele trei categorii de adnotări de conţinut, în Figura 3.9 am ilustrat un exemplu concret de adnotare sintactică, simbolică şi respectiv semantică în cazul unei secvenţe de fotbal (din motive de vizualizare, secvenţa este reprezentată prin ilustrarea a câtorva imagini reprezentative).

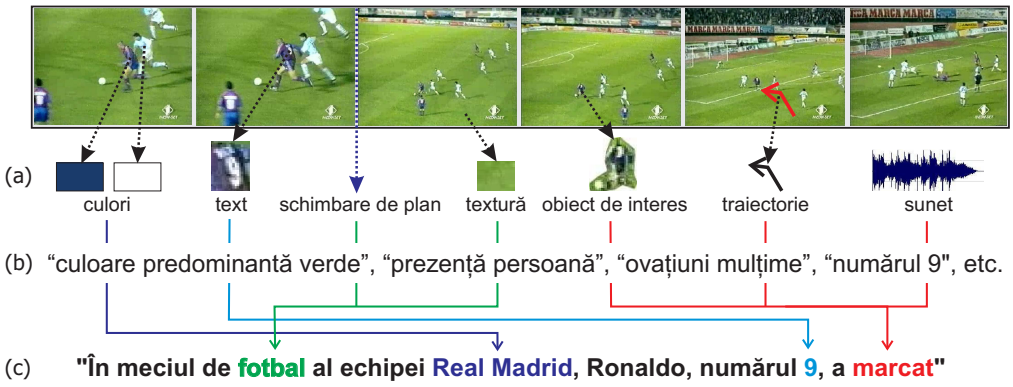


Figura 3.9: Exemplu de descriere sintactică (a), simbolică (b) şi semantică (c) în cazul unei secvenţe de imagini (axa orizontală reprezintă axa temporală).

Astfel, pornind de la informaţia video (imagine-sunet), adnotarea sintactică va fi capabilă să furnizeze doar informaţii relative la scenă şi la pro-

prietăţile acesteia, precum culoare, prezenţă text, textură, traiectoria obiectelor în mişcare, ritmul de desfăşurare al acţiunii sau detecţie zgomot audio specific mulţime.

Folosind aceste informaţii se poate obţine o descriere simbolică de nivel semantic intermediar al conţinutului video precum detecţia culorii predominante ce corespunde gazonului, detecţia unei persoane în mişcare, detecţia ovaţiunilor mulţimii specifice unui gol, detecţia tricoului cu numărul 9 şi aşa mai departe. Aceste informaţii, nu sunt simple date numerice dar totuşi nu furnizează o întelegere semantică a conţinutului secvenţei.

O adnotare semantică va da sens acestor informaţii într-un mod unitar, de exemplu textura verde va indica că este vorba despere un meci de fotbal, culorile jucătorilor (obiecte în mişcare) vor dezvălui echipele, recunoaşterea numerelor de pe tricou va identifica jucătorii, segmentarea obiectului de interes, urmărirea acestuia şi prezenţa zgomotului specific mulţimii vor indica marcarea golului. Astfel că sistemul va "înţelege" sensul acţiunii secvenţei şi anume că este vorba despre un meci de fotbal al echipei Real Madrid în care jucătorul cu numărul 9, Ronaldo, marchează.

CAPITOLUL 4

Fuziunea datelor

În cele mai multe dintre cazuri, pentru reprezentarea conţinutului multimedia este necesară combinarea mai multor tipuri de descriptori. De exemplu, conţinutul unei secvenţe de imagini poate fi reprezentat atât pe baza structurii temporale, cât şi folosind descriptori de mişcare, descriptori audio şi aşa mai departe. Strategiile de fuziune a datelor se bazează pe ipoteza conform căreia o decizie obţinută pe baza mai multor descriptori poate oferi performanţe superioare unei decizii bazate pe un singur tip de descriptor.

Astfel, se pune problema găsirii unei modalităţi de agregare (fuziune) a acestor date, formând în general un nou descriptor ce sintetizează cât mai bine puterea discriminatorie a descriptorilor individuali.

Cu alte cuvinte, ideal, noul descriptor trebuie să păstreze acele proprietăţi distincte ale descriptorilor individuali (de exemplu informaţia audio descrie proprietăţi diferite faţă de informaţia structurală) şi să elimine informaţiile redundante (similare), exploatând cât mai bine complementaritatea acestora în reprezentarea informaţiei. În general există două tipuri de abordări ale problemei fuziunii datelor, tehnici de tip "*early fusion*" şi respectiv "*late fusion*" [Snoek 05].

4.1 Metode de tip "early fusion"

Tehnicile de tip "early fusion" realizează agregarea datelor "timpuriu" în lanţul de prelucrare, înainte de a fi folosite la indexare sau în alte procese de analiză. Fuziunea datelor are loc în spaţiul de caracteristici (vezi Secţiunea

2.1) şi constă practic în concatenarea propriu-zisă a tuturor descriptorilor fără a ţine cont de redundanţa acestora.

De exemplu, dacă obiectul multimedia X este descris de descriptorii de conţinut $desc_1 = \{a_1, a_2, ..., a_n\}$, $desc_2 = \{b_1, b_2, ..., b_m\}$ şi respectiv $desc_3 = \{c_1, c_2, ..., c_l\}$, unde a, b şi c reprezintă valorile atributelor acestora, descriptorul agregat este dat de concatenarea valorilor şi anume $desc_{e-f} = \{a_1, ..., a_n, b_1, ..., b_m, c_1, ..., c_l\}$. Acesta defineşte astfel un nou spaţiu de caracteristici $(n + m + l)$-dimensional.

O problemă care apare o reprezintă necesitatea normalizării valorilor datelor într-un anumit interval comun. Descriptori diferiţi tind să aibă intervale de variaţie diferite ale valorilor, de la normalizări diferite, de exemplu valori între $[0; 1]$ sau $[a; b]$ (unde a şi b sunt două valori cunoscute) până la intervale de valori variabile şi care depind de tipul datelor.

Dintre tehnicile de normalizare cel mai frecvent folosite putem enumera normalizarea min-max:

$$a_i = \frac{a_i - min\{a_i\}}{max\{a_i\} - min\{a_i\}} \tag{4.1}$$

unde a_i sunt atributele descriptorului, $i = 1, ..., n$ cu n numărul de valori ale acestuia, $min\{a_i\}$ şi $max\{a_i\}$ reprezintă operatorii ce returnează valoarea minimă şi respectiv maximă a tuturor valorilor descriptorilor (pentru toate obiectele multimedia considerate) pentru atributul a_i. Calculată în acest fel, normalizarea min-max asigură o normalizare a valorilor în intervalul $[0; 1]$.

Normalizarea z-score se foloseşte de calculul abaterii pătratice medii:

$$a_i = \frac{a_i - medie\{a_i\}}{\sigma\{a_i\}} \tag{4.2}$$

unde ca şi în cazul anterior, operatorii $medie\{a_i\}$ şi $\sigma\{a_i\}$ returnează valoarea medie şi respectiv abaterea pătratică medie a tuturor valorilor descriptorilor pentru atributul a_i. În acest caz normalizarea se realizează pe o distribuţie de medie zero şi dispersie unu.

O altă abordare constă în calculul statisticii mediane:

$$a_i = \frac{a_i - median\{a_i\}}{median\{|a_i - median\{a_i\}|\}} \tag{4.3}$$

unde operatorul $median\{a_i\}$ returnează statistica mediană[1] a mulţimii tuturor valorilor descriptorilor pentru atributul a_i iar opertorul $|.|$ returnează valoarea absolută.

[1] valoarea mediană a unei mulţimi se obţine prin ordonarea valorilor acesteia în ordine crescătoare si alegerea valorii de mijloc.

Dacă ordinul intervalului de variaţie al valorilor descriptorului diferă foarte mult, ca de exemplu printr-un ordin de mărime logaritmic, $[0; 1]$ comparativ cu $[0; 1000]$, normalizarea se poate realiza folosind scalarea zecimală:

$$a_i = \frac{a_i}{10^n}, \quad n = log_{10}(max\{a_i\}) \tag{4.4}$$

În cazul în care nu se cunoaşte intervalul de variaţie al valorilor descriptorului se poate opta pentru o normalizare folosind funcţii duble sigmoide:

$$a_i = \left[1 + exp\left(-2 \cdot \frac{a_i - t}{r}\right)\right]^{-1} \tag{4.5}$$

unde t este de regulă valoarea medie a distribuţiei valorilor descriptorului iar $r = r_1$ dacă $a_i < t$ şi $r = r_2$ în caz contrar. Constantele r_1 şi r_2 reprezintă valorile unor intervale alese la dreapta şi respectiv stânga valorii lui t. Aceste aspecte sunt ilustrate în Figura 4.1.

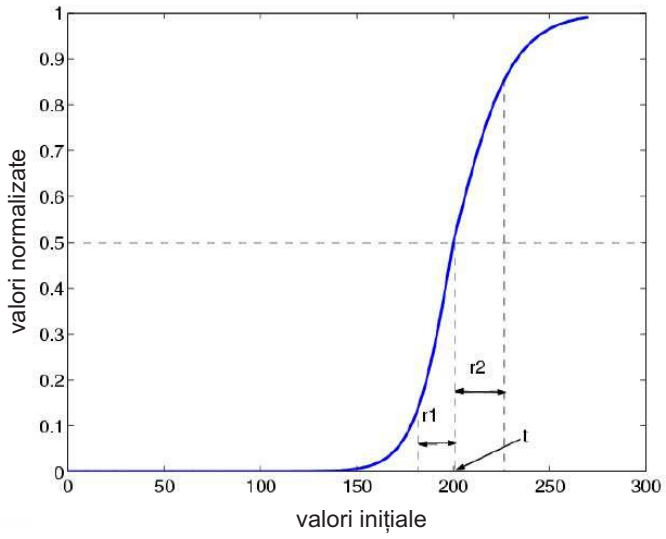


Figura 4.1: Exemplu de normalizare folosind funcţii dublu sigmoide (axa oX corespunde valorilor iniţiale iar axa oY valorilor normalizate).

Principalul dezavantaj al tehnicilor de tip "early fusion" este dat de dimensionalitatea datelor, descriptorul obţinut prin fuziune având ca dimensiune suma dimensiunilor descriptorilor individuali, ceea ce conduce la un număr semnificativ de valori (un astfel de descriptor agregat în cazul video poate avea uzual zeci de mii de componente).

Cu cât dimensiunea datelor este mai ridicată cu atât este mai probabil ca puterea discriminatorie să scadă deoarece datele similare tind să se disperseze în spaţiul de caracteristici ceea ce face dificilă separarea acestora (şi astfel indexarea). De asemenea, folosind concatenarea descriptorilor nu se poate controla contribuţia pe care o are fiecare descriptor individual asupra sistemului. Descriptorii de dimensiune mai mare vor tinde să aibă pondere principală în reprezentarea conţinutului raportat la descriptorii cu un număr redus de valori (de exemplu descriptorii care conţin o singură valoare).

4.2 Metode de tip "late fusion"

Pe de altă parte, tehnicile de tip "late fusion" realizează fuziunea "târziu" în lanţul de prelucrare bazându-se pe exploatarea individuală a puterii discriminatorii a fiecărui descriptor sau modalităţi în parte.

Să considerăm pentru exemplificare un sistem de indexare după conţinut bazat pe clasificarea datelor. Tehnicile de clasificare sunt tehnici de învăţare asistată de calculator ("machine learning"). Problema pe care o rezolvă poate fi formulată în felul următor: având la dispoziţie un set necunoscut de date se doreşte realizarea unei partiţionări a acestora în clase de similaritate (etichetarea acestora ca aparţinând unei anumite categorii). Pentru aceasta, sistemul poate dispune de o serie de exemple de partiţii, numite şi date de antrenare - date pentru care se cunoaşte apartenenţa la clase şi pentru care problema clasificării este deja soluţionată (de regulă de către un expert). Pe baza datelor de antrenare, clasificatorul învaţă mecanismul de asociere în clase urmând să-l aplice ulterior datelor noi necunoscute (ne-etichetate) [Witten 05]. Principiul este ilustrat schematic în Figura 4.2.

În contextul indexării după conţinut, tehnicile de clasificare transpun problema căutării într-o problemă inversă de partiţionare a bazei de date în funcţie de conţinutul căutat. Problema indexării se transpune astfel într-o problemă de partiţionare adecvată a datelor în categoriile căutate de utilizator. De exemplu, dacă se doreşte căutarea unui anumit gen video, baza de date va fi clasificată după diferite clase de gen (de exemplu film, muzică, ştiri), dacă se doreşte găsirea unui anumit obiect într-o bază de imagini, acestea vor fi clasificate în clase de obiecte (de exemplu "minge", "maşină", "casă"). Procesul de clasificare se realizează pe baza reprezentării datelor cu descriptori de conţinut.

În momentul căutării datelor, cererea de căutare a utilizatorului ("query") va fi asociată uneia dintre clasele determinate anterior, rezultatele căutării fiind acele documente ce au fost etichetate ca aparţinând acestei clase. Ca şi în cazul mecanismului de indexare clasic (vezi Secţiunea 2) datele vor fi

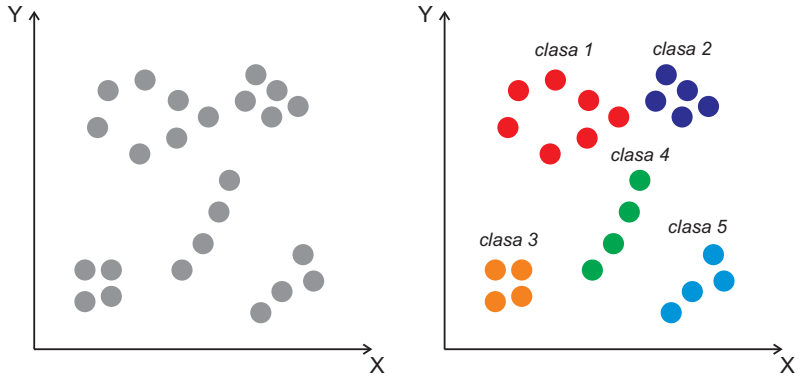


Figura 4.2: Principiul clasificării datelor: (a) datele de intrare reprezentate în spaţiul de caracteristici, (b) repartiţia în clase obţinută în urma clasificării (obiectele din aceeaşi clasă sunt reprezentate cu aceeaşi culoare).

returnate în ordinea descrescătoare a relevanţei. Pentru ca acest lucru să fie posibil, clasificatorul în locul unei decizii binare de apartenenţă sau non-apartenenţă va furniza un grad de relevanţă, de regulă o valoare reală în intervalul $[0; 1]$, unde 1 reprezintă apartenenţa sigură la clasă, iar 0 cazul contrar. Astfel, rezultatele sunt returnate utilizatorului în ordinea descrescătoare gradului de relevanţă furnizat de clasificator pentru clasa ce aparţine căutării. Acest mecanism este exemplificat în Figura 4.3 în contextul căutării după gen a documentelor video.

Revenind la problematica fuziunii datelor, fuziunea de tip "late fusion" se realizează în acest caz prin fuzionarea rezultatelor clasificărilor obţinute independent pentru fiecare tip de descriptor sau modalitate, cât şi pentru tipuri de clasificatori diferiţi. În acest fel, agregarea datelor nu este realizată la nivel de descriptor ci la nivelul gradului de relevanţă atribuit de fiecare clasificator descriptorilor, beneficiind de puterea discriminatorie a fiecărui descriptor în parte. Dintre tehnicile de tip "late fusion" cel mai frecvent folosite putem enumera:

- **fuziunea paralelă**: presupune rularea aceluiaşi sistem în paralel pentru descriptori şi tipuri de clasificatori diferiţi. Agregarea finală a rezultatelor se realizează pe baza agregării rezultatelor obţinute individual (vezi Figura 4.4);

- **fuziunea serială**: presupune rularea în cascadă a sistemelor, fiecare ieşire a unui clasificator fiind folosită la intrarea unui alt clasificator ca de exemplu pentru clasificarea datelor ce au fost clasificate eronat

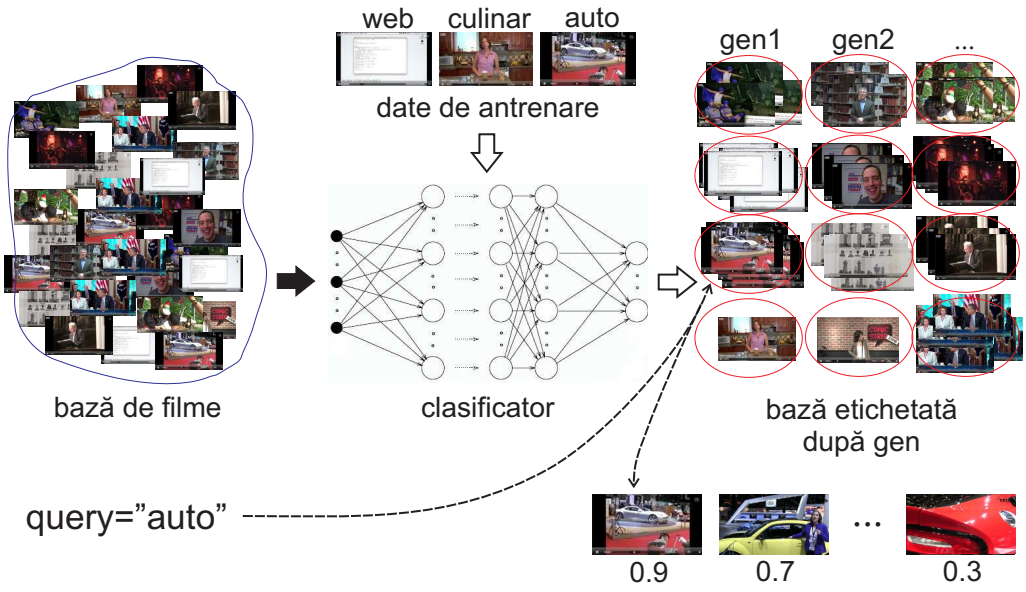


Figura 4.3: Exemplu de sistem de clasificare folosit pentru căutarea după gen a secvenţelor video. Clasificatorul este mai întâi antrenat folosind un set redus de exemple şi un set predefinit de genuri urmând să catalogheze automat baza video necunoscută. Cererea de căutare primeşte ca rezultat secvenţele ce au fost atribuite clasei căutate în ordinea descrescătoare a gradului de relevanţă furnizat de clasificator (sursă imagini blip.tv).

de sistemul anterior. Fiecare dintre sisteme rulează pentru descriptori şi clasificatori diferiţi. Principiul este inspirat de tehnicile de tip "boosting" în care mai mulţi clasificatori "slabi" (cu preformanţe reduse) sunt combinaţi pentru a obţine un clasificator cu performanţe ridicate (vezi AdaBoost [Witten 05]);

- **fuziunea ierarhică**: sistemele sunt organizate ierarhic, fie de tip "bottom-up" în care mai mulţi clasificatori converg către un clasificator final, sau de tip "top-down" unde în funcţie de rezultatele unui clasificator iniţial, deciziile se separă ierarhic pe mai multe niveluri de clasificatori. Acest mod de reprezentare este similar arborilor de decizie (vezi Random Forest sau Random Tree [Witten 05]).

- **fuziunea mixtă**: constă în combinarea mai multor modalităţi de fuzionare din categoriile enumerate anterior.

În continuare vom detalia modul de luare al deciziei în cazul fuzionării paralele. Acesta este ilustrat în Figura 4.4. Având la dispoziţie N clasi-

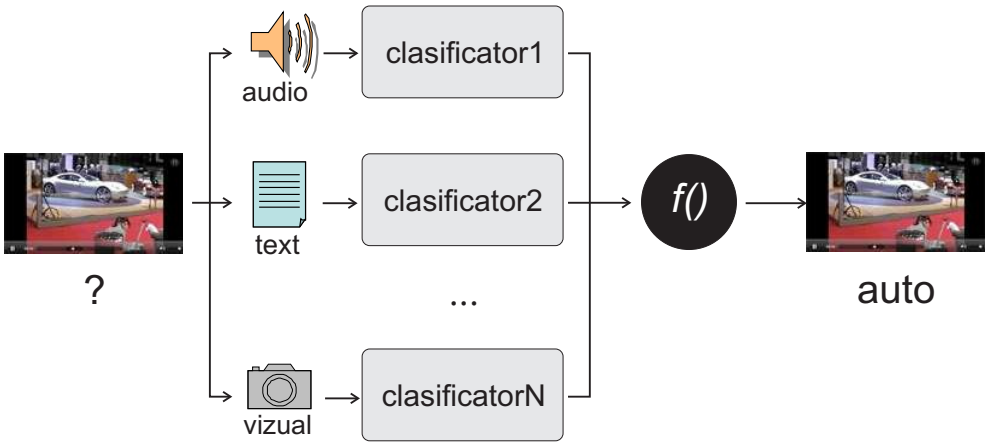


Figura 4.4: Principiul fuzionării de tip "late fusion" paralel. Catalogarea datelor de intrare se realizează pe baza unei funcţii de agregare, $f(.)$, a ieşirilor mai multor tipuri de clasificatori antrenaţi folosind descriptori diferiţi.

ficatori ce sunt antrenaţi folosind descriptori de conţinut diferiţi, fuzionarea de tip "late fusion" a descriptorilor presupune determinarea unei funcţii care combină gradele de relevanţă furnizate de fiecare clasificator în parte, $f(x_1, ..., x_N)$, unde x_i reprezintă gradul de relevanţă atribuit de clasificatorul i datelor de intrare. Acestea reprezintă probabilităţile de apartenenţă la clasele considerate, $x_i = \{p_{i,c_1}, p_{i,c_2}, ..., p_{i,c_M}\}$ unde $c_1, ..., c_M$ reprezintă clasele considerate (de exemplu genurile video în Figura 4.3) iar $p_{i,c}$ reprezintă probabilitatea ca datele să fie atribuite ca aparţinând clasei c.

În mod natural, fiecare clasificator va tinde să furnizeze grade de apartenenţă diferite fiind antrenat pentru descriptori diferiţi. Funcţia $f(.)$ trebuie determinată în aşa fel încât rezultatele obţinute de clasificatorul agregat să fie cât mai bune şi superioare fiecărui clasificator individual. Agregarea se va realiza pentru gradele de relevanţă ale fiecărei clase în parte.

O modalitate de definire a lui $f(.)$ o reprezintă combinaţia liniară a gradelor de relevanţă şi anume:

$$f_{CombMean}(d, c_j) = \sum_{i=1}^{N} \alpha_i \cdot p_{i,c_j} \tag{4.6}$$

unde d reprezintă documentul curent, p_{i,c_j} reprezintă probabilitatea de apartenenţă la clasa c_j, $j = 1, ..., M$ cu M numărul de clase considerate, atribuită de clasificatorul i iar α_i reprezintă un set de ponderi. Un caz particular îl reprezintă considerarea de ponderi egale ceea ce conduce la însumarea gradelor de relevanţă pentru fiecare clasă.

Un alt exemplu este atribuirea unei ponderi superioare acelor date care sunt mai probabile să fie relevante pentru o clasă, astfel:

$$f_{CombMNZ}(d, c_j) = F(d)^{\gamma} \cdot \sum_{i=1}^{N} \alpha_i \cdot p_{i,c_j} \tag{4.7}$$

unde $F(d)$ reprezintă numărul de clasificatori pentru care documentul d apare în primele k documente din punct de vedere al valorii de relevanţă (k este o constantă stabilită a priori) iar $\gamma \in [0, 1]$ este un parametru de control.

Noile valori de relevanţă obţinute în urma agregări sunt folosite mai departe pentru indexarea datelor în mod similar în care acestea erau folosite în cazul considerării unui singur clasificator.

Comparate cu abordările de tip "early fusion", tehnicile de tip "late fusion" sunt mai avantajoase din punct de vedere computaţional deoarece agregarea se face folosind dimensiunea iniţială a descriptorilor. Este mai eficientă clasificarea unor descriptori de dimensiuni reduse şi agregarea rezultatelor decât clasificarea unui descriptor agregat de dimensiuni semnificativ mai mari. Principalul dezavantaj al acestor metode este totuşi dat de pierderea eventualei corelaţii dintre descriptori ce se obţine în cazul concatenării acestora şi care poate furniza un nivel de discriminare superior folosirii individuale a acestora.

În ciuda diferenţelor dintre cele două abordări, "early fusion" şi respectiv "late fusion", nu există o metodă preferenţială în defavoarea celeilalte, ambele abordări dovedindu-se eficiente în contexte diferite. Astfel că tehnica de fuziune a datelor rămâne dependentă de aplicaţie [Lan 12].

CAPITOLUL 5

Conceptul de similaritate a datelor

Aşa cum am prezentat în Secţiunea 2.3, în procesul de căutare după conţinut a datelor, descrierea eficientă a conţinutului nu este suficientă pentru a asigura indexarea acestora în baza de date. La fel de importantă este definirea conceptului de similaritate (sau opus, disimilaritate) dintre date sau dintre descriptorii acestora.

Practic identificarea rezultatelor căutării se realizează prin localizarea datelor ce sunt "similare" până la un anumit nivel cu cererea de căutare ("query"). Cu alte cuvinte este necesară definirea unei funcţii, $S(O_1, O_2)$, capabilă să evalueze în ce măsură două obiecte multimedia, O_1 şi O_2, arată sau sună în mod similar, în ce măsură au o structură similară sau în ce măsură conduc la aceeaşi percepţie sau interpretare a conţinutului [Worring 03].

În general, evaluarea similarităţii dintre date se poate realiza fie la nivel de *descriptori*, la nivel de *structură* ("layout") sau la nivel *semantic*, fie folosind combinaţii ale acestora.

5.1 Similaritatea descriptorilor

În acest caz, similaritatea datelor este evaluată numeric folosind valorile descriptorilor de conţinut aferente acestora iar funcţia $S()$ este de regulă o măsură de distanţă (metrică). Datele vor fi considerate similare în măsura în care valoarea distanţei dintre descriptorii acestora este minimă.

În cele ce urmează vom face o trecere în revistă a diverselor metrici folosite în domeniul căutării informaţiei. Marea partea dintre acestea sunt în mod

natural inspirate din matematică [Deza 06].

Una dintre abordările clasice este folosirea distanţei Minkovski, ce este definită ca:

$$S_{Mink}(A_{O_1}, A_{O_2}) = \sqrt[r]{\sum_{i=1}^{n} [A_{O_1}(i) - A_{O_2}(i)]^r} \tag{5.1}$$

unde $A_O(i)$ reprezintă valoarea de indice i a descriptorului aferent obiectului multimedia O, cu $i = 1, ..., n$ elemente (de regulă descriptorii de conţinut sunt vectori n-dimensionali de valori, vezi şi Secţiunea 2.1).

În cazul în care considerăm parametrul $r = 1$ obţinem norma L1 sau distanţa Manhattan:

$$S_{Manh}(A_{O_1}, A_{O_2}) = \sum_{i=1}^{n} |A_{O_1}(i) - A_{O_2}(i)| \tag{5.2}$$

unde operatorul $|.|$ reprezintă valoarea absolută.

Pentru $r = 2$ obţinem mai departe norma L2 cunoscută sub numele de distanţa Euclidiană:

$$S_{Euclid}(A_{O_1}, A_{O_2}) = \sqrt{\sum_{i=1}^{n} [A_{O_1}(i) - A_{O_2}(i)]^2} \tag{5.3}$$

În cazul în care nu toate elementele descriptorului au aceeaşi importanţă, distanţa dintre fiecare pereche de valori poate fi ponderată diferit obţinând astfel distanţa Euclidiană ponderată:

$$S_{wEuclid}(A_{O_1}, A_{O_2}) = \sqrt{\sum_{i=1}^{n} w_i \cdot [A_{O_1}(i) - A_{O_2}(i)]^2} \tag{5.4}$$

unde w_i, cu $i = 1, ..., n$ reprezintă ponderile fiecărei valori.

O altă măsură de distanţă ce este folosită de regulă când descriptorii de conţinut sunt reprezentaţi sub formă de histograme (de exemplu histograma color a unei imagini) o constituie intersecţia histogramei. Aceasta este de fapt o măsură a disimilarităţii şi este definită ca suma minimelor valorilor histogramelor:

$$S_{inter}(h_{O_1}, h_{O_2}) = \sum_{i=1}^{n} min\{h_{O_1}(i), h_{O_2}(i)\} \tag{5.5}$$

unde $h_O(i)$ cu $i = 1, ..., n$ reprezintă histograma color a obiectului multimedia O iar operatorul $min\{.\}$ returnează valoarea minimă a unui set de elemente.

Tot în cazul evaluări diferenţelor dintre histograme şi în special dintre histogramele color ale imaginilor, în cazul folosirii distanţelor clasice, este foarte probabil ca pentru distribuţii ale unei aceleiaşi nuanţe (de exemplu roşu deschis şi roşu) să obţinem valori semnificative ale distanţei, de ordin de măsură similar ca pentru distanţa faţă de o distribuţie a unei nuanţe complet diferite (de exemplu albastru), în ciuda faptului că diferenţele în primul caz ar trebui să fie reduse, culorile fiind asemănătoare. O distanţă care tinde să contracareze acest efect este distanţa pătratică dintre histograme:

$$S_{hist2}(h_{O_1}, h_{O_2}) = \sqrt{(h_{O_1} - h_{O_2})^T \cdot A \cdot (h_{O_1} - h_{O_2})} \tag{5.6}$$

unde h_O reprezintă vectorul histogramă cu n elemente, T reprezintă transpusa unei matrice iar $A = [a_{i,j}]$, $i, j = 1, ..., n$, reprezintă o matrice pătratică de valori ce indică corelaţia dintre elementele histogramelor de indici i cu cele de indice j (de regulă A este simetrică şi are elementele de pe diagonala principală egale cu 1).

Alte măsuri de distanţă frecvent folosite sunt distanţa Canberra:

$$S_{Canb}(A_{O_1}, A_{O_2}) = \sum_{i=1}^{n} \frac{|A_{O_1}(i) - A_{O_2}(i)|}{|A_{O_1}(i)| + |A_{O_2}(i)|} \tag{5.7}$$

distanţa Bray-Curtis:

$$S_{B-C}(A_{O_1}, A_{O_2}) = \frac{\sum_{i=1}^{n} |A_{O_1}(i) - A_{O_2}(i)|}{\sum_{i=1}^{n} [A_{O_1}(i) + A_{O_2}(i)]} \tag{5.8}$$

distanţa SquaredChord:

$$S_{SChord}(A_{O_1}, A_{O_2}) = \sum_{i=1}^{n} \left[\sqrt{A_{O_1}(i)} - \sqrt{A_{O_2}(i)}\right]^2 \tag{5.9}$$

distanţa Lorentzian, Soergel, Czekanowski, Motyka, Ruzicka, Tanimoto, Wave-Hadges, Clark, Person şi aşa mai departe. Pentru mai multe detalii cititorul se poate raporta la [Deza 06].

O abordare diferită este distanţa Bhattacharyya ce măsoară similaritatea a două distribuţii de probabilitate. În cazul în care descriptorii sunt consideraţi a avea o distibuţie normală Gaussiană, distanţa poate fi scrisă ca fiind:

$$S_{Bhatta}(A_{O_1}, A_{O_2}) = \frac{1}{8} \cdot \left(\mu_{A_{O_1}} - \mu_{A_{O_2}}\right)^T \cdot (\Sigma_{O_1,O_2})^{-1} \cdot \left(\mu_{A_{O_1}} - \mu_{A_{O_2}}\right) + \frac{1}{2} \cdot ln\left(\frac{det(\Sigma_{O_1,O_2})}{\sqrt{det(\Sigma_{O_1}) \cdot det(\Sigma_{O_2})}}\right) \tag{5.10}$$

unde μ_{A_O} reprezintă vectorul medie al distribuţiei de probabilitate a descriptorului A_O, Σ_O reprezintă matricea de covarianţă a distribuţiei lui A_O, Σ_{O_1,O_2} reprezintă media aritmetică a matricelor de covarianţă pentru distribuţiile lui A_{O_1} şi A_{O_2} (vezi şi [Ciuc 05]), T reprezintă transpusa unei matrice iar operatorul $det(.)$ returnează determinantul unei matrice.

O altă perspectivă o constituie reprezentarea datelor sub formă de mulţimi. Distanţa Hausdorff evaluează gradul de apropiere a două submulţimi într-un anumit spaţiu şi folosind o anumită metrică, astfel:

$$S_{Haus}(A_{O_1}, A_{O_2}) = max\{\sup_i \inf_j d(A_{O_1}(i), A_{O_2}(j)), \sup_j \inf_i d(A_{O_1}(i), A_{O_2}(j))\} \quad (5.11)$$

unde $i, j = 1, ..., n$, inf şi sup reprezintă infimum şi respectiv supremum al unei mulţimi (de regulă valoarea minimă şi respectiv maximă), $d(.)$ reprezintă o anumită metrică (de exemplu norma L1) iar $max\{.\}$ returnează valoarea maximă a unei mulţimi. În acest caz, valorile descriptorilor pot fi văzute din perspectiva elementelor unei mulţimi.

Un alt caz interesant este distanţa cosinus. Să presupunem că descriptorii de conţinut sunt vectori de caractere iar datele ce trebuiesc comparate sunt documente textuale, atunci similaritatea dintre acestea poate fi evaluată folosind produsul scalar:

$$S(A_{O_1}, A_{O_2}) = \sum_{i=1}^{n} A_{O_1}(i) \cdot A_{O_2}(i) \quad (5.12)$$

Acum dacă descriptorii textuali sunt reprezentaţi sub formă de histograme ale căror valori indică numărul de apariţii al unui anumit cuvânt în document (eventual ponderat de un factor de importanţă - cuvintele sunt alese pentru un dicţionar predefinit; vezi TF-IDF în Secţiunea 3.3) atunci similaritatea se reduce la o înmulţire a valorilor histogramelor pentru cele două documente. Astfel, atunci când un cuvânt apare frecvent în cele două documente, contribuţia acestuia la produs va fi semnificativă.

Problema care apare este faptul că documentele mari vor conţine mai multe cuvinte şi vor tinde să devină mai similare decât documentele ce conţin mai puţin text. Astfel că în practică descriptorii sunt normalizaţi la dimensiunea acestora $||A_O||^2 = \sum_{i=1}^{n} A_O^2(i)$ ceea ce conduce la formularea distanţei cosinus astfel:

$$S_{cos}(A_{O_1}, A_{O_2}) = \frac{A_{O_1} \cdot A_{O_2}}{||A_{O_1}|| \cdot ||A_{O_2}||} \quad (5.13)$$

unde $\cdot$ reprezinta produsul scalar (denumirea de cosinus vine de la faptul că distanţa este practic cosinusul unghiului celor doi vectori normalizaţi).

În cazul comparării de obiecte, de exemplu prin intermediul a două imagini binare (în care obiectul are valoarea 1 şi fundalul 0) se poate folosi distanţa Baddeley definită în felul următor:

$$S_{Badd}(I_{O_1}, I_{O_2}) = \left[\frac{1}{M \cdot N} \sum_{p \in S} |d_{I_{O_1}}(p) - d_{I_{O_2}}(p)|^q \right]^{1/q} \tag{5.14}$$

unde I_O reprezintă o imagine binară, $M \cdot N$ reprezintă numărul total de pixeli din setul S, $d_{I_O}(p)$ reprezintă o anumită metrică de distanţă de la punctul p la cel mai apropiat punct al obiectului conţinut în imaginea I_O iar q este exponentul (de regulă se consideră $q = 2$). Definită în acest fel, distanţa Baddeley oferă un anumit grad de invarianţă la translaţia obiectelor şi modificarea factorului de scală.

O problemă aparte o ridică compararea descriptorilor de dimensiuni diferite, ca de exemplu histogramele color a două imagini cu palete de culoare diferite (binii histogramei şi numărul acestora sunt diferite). O soluţie în acest sens este propusă de distanţa Earth Mover's Distance (EMD). Aceasta se bazează pe evaluarea costului minim aferent transformării unuia dintre descriptori în celălalt şi este formulată ca o problemă de optimizare. EMD este definită în felul următor:

$$S_{EMD}(A_{O_1}, A_{O_2}) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{i,j} \cdot f_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j}} \tag{5.15}$$

unde cei doi descriptori A_{O_1} şi respectiv A_{O_2} au dimensiuni diferite, m şi respectiv n, $d_{i,j}$ reprezintă distanţa dintre valorile $A_{O_1}(i)$ şi respectiv $A_{O_2}(j)$ iar $f_{i,j}$ este o funcţie de cost ce reprezintă deplasarea între $A_{O_1}(i)$ şi $A_{O_2}(j)$ determinată ca minimizând valoarea costului total $\sum_{i=1}^{m} \sum_{j=1}^{n} d_{i,j} \cdot f_{i,j}$ cu o serie de constrângeri [Rubner 00].

O altă categorie de distanţe sunt cele inspirate din teoria informaţiei a lui Shannon, precum divergenţele Kullback–Leibler:

$$S_{K-L}(A_{O_1}, A_{O_2}) = \sum_{i=1}^{n} A_{O_1}(i) \cdot \ln \frac{A_{O_1}(i)}{A_{O_2}(i)} \tag{5.16}$$

sau divergenţa Jeffrey:

$$S_{Jeff}(A_{O_1}, A_{O_2}) = \sum_{i=1}^{n} [A_{O_1}(i) - A_{O_2}(i)] \cdot [ln(A_{O_1}(i)) - ln(A_{O_2}(i))] \tag{5.17}$$

Acestea sunt aplicate cu precădere la compararea descriptorilor specifici datelor audio, unde este relevantă distribuţia statistică a valorilor acestora.

Pentru a ilustra importanţa alegerii adecvate a măsurii de distanţă, în Figura 5.1 am prezentat rezultatele obţinute pentru o căutare de imagini cu "relevance feedback" (vezi Secţiunea 2.4) şi folosind metrici şi descriptori de conţinut diferiţi [Mironică 12b]. Graficele ilustrează performanţa căutării pe baza valorii MAP (Mean Average Precision, vezi Secţiune 8; reprezentată pe axa oY - valoarea maximă este 1 ce indică o performanţă de 100%) raportată la metrica folosită (axa oX). Pentru descrierea conţinutului imaginilor au fost folosiţi descriptori de trăsături de tip SIFT şi SURF (vezi Secţiune 3.1). Testele au fost efectuate pe două baze de imagini, baza Microsoft Object Class Recognition[1] (puncte roşii) şi respectiv Caltech-101[2] (puncte albastre).

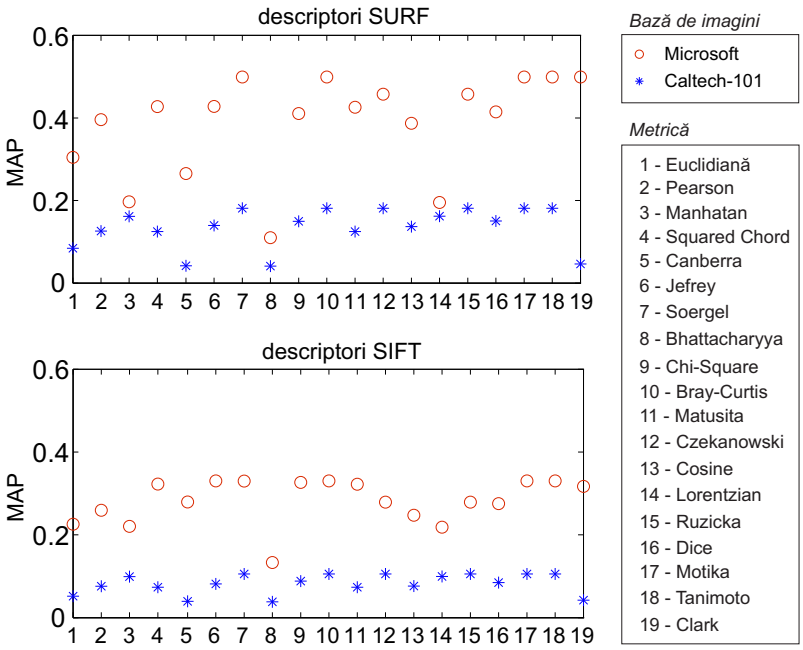


Figura 5.1: Exemplu de influenţă a metricii asupra performanţelor căutării de imagini [Mironică 12b] (MAP reprezintă Mean Average Precision - valoarea maximă 1 corespunde unei performanţe de 100%).

Se poate observa faptul că în funcţie de metrică, performanţele sistemului variază semnificativ, de exemplu pentru baza Microsoft valorile MAP variază de la 10% la 50% pentru descriptorii SURF şi de la 10% la 38% pentru SIFT. Pe lângă alegerea adecvată a descriptorilor (se observă şi în figură faptul că

[1]Microsoft Object Class Recognition `http://research.microsoft.com/en-us/projects/objectclassrecognition`.

[2]Caltech-101 `http://www.vision.caltech.edu/Image_Datasets/Caltech101`.

descriptorul SURF este mai performant în contextul sistemului prezentat), alegerea adecvată a metricii joacă un rol cel puţin la fel de important.

5.2 Similaritatea la nivel de structură

Aceasta presupune evaluarea gradului de similaritate a două obiecte multimedia, O_1 şi O_2, din punct de vedere al structurii acestora, ca de exemplu modul de aranjare spaţială a obiectelor în imagini, modul de structurare al unei paginii de text, structura temporală a unui document video. O modalitate eficientă de caracterizare a structurii este prin intermediul descrierii acesteia cu şiruri de caractere [Worring 03].

Să considerăm în continuare exemplul datelor video. Un document video, din punct de vedere structural, este constituit ca o înşiruire de plane video separate de tranziţii (vezi Secţiunea 3.1). Informaţia structurală poate consta în descrierea acestei structuri. Documentul video poate fi reprezentat ca un şir de caractere de genul "*scswsdcs*", unde s reprezintă un plan video ("shot"), c reprezintă o tranziţie de tip "cut", w reprezintă o tranziţie graduală de tip "wipe" iar d reprezintă un "dissolves". Informaţia temporală este dată de ordinea simbolurilor în şir, astfel acest document video începe cu un plan urmat de un "cut", un plan video, o tranziţie "dissolves" şi aşa mai departe.

Pentru a compara similaritatea descriptorilor astfel obţinuţi o variantă eficientă o reprezintă folosirea distanţei de editare ("edit distance"), ce foloseşte un concept similar distanţei Earth Mover's Distance (EMD) descrisă anterior. Având la dispoziţie descriptorii structurali de conţinut ai celor două obiecte multimedia, $A_{O_1} = \{a_{1,1}, a_{1,2}, ..., a_{1,n}\}$ şi respectiv $A_{O_2} = \{a_{2,1}, a_{2,2}, ..., a_{2,m}\}$, unde n şi m reprezintă numărul de caractere, un alfabet Σ ce descrie simbolurile posibile (valorile lui a), un set E de operaţii de editare şi costurile aferente acestora, distanţa de editare dintre A_{O_1} şi A_{O_2} reprezintă costul minim de transformare a şirului A_{O_1} în şirul A_{O_2} pe baza operaţiilor din E.

În cazul a două secvenţe video, să presupunem că descriptorii acestora sunt $V_{O_1} = \{scswsdcs\}$ şi $V_{O_2} = \{sdswscscscs\}$, mulţimea $\Sigma = \{c, w, d, s\}$ iar operaţiile de editare posibile sunt $E = \{$*inserare,ştergere,înlocuire*$\}$ iar costurile aferente acestora sunt egale. Operaţiile necesare pentru a transforma pe V_{O_1} în V_{O_2} constau în două "*înlocuiri*" ale lui c cu d şi respectiv două operaţii de "*inserare*" pentru adăugarea lui c şi s la sfârşit. Astfel distanţa de editare dintre cei doi descriptori este în acest caz 4.

5.3 Similaritatea semantică

Aşa cum am prezentat şi în Secţiunea 3.4, tendinţa actuală a sistemelor de indexare este aceea de a determina descriptori de conţinut ce oferă un nivel de înţelegere al conţinutului cât mai apropiat de nivelul de percepţie uman. Acest lucru se realizează în principal prin identificarea a ceea ce numim "concepte". Un concept este practic o reprezentare textuală a entităţilor reprezentate de obiectul multimedia, exemple de concepte în cazul imaginilor sau a documentelor video fiind "cer", "maşină", "persoană", "casă", şi aşa mai departe [Over 12].

Reprezentarea conceptelor poate fi realizată fie prin adnotarea manuală a acestora de către utilizator, fie folosind tehnici automate de adnotare sau folosind informaţii derivate din ontologii. Ontologia constituie un mod formal de reprezentare a cunoaşterii sub forma unui set de concepte dintr-un domeniu şi a relaţiilor dintre acestea folosind următoarele componente: obiecte sau instanţe, clase (mulţimi, colecţii, concepte), atribute (proprietăţi, trăsături, parametri ai obiectelor şi claselor), relaţii (descriu modul în care clasele şi instanţele sunt relaţionate), restricţii, reguli (afirmaţii de tip dacă-atunci ("if-then") sau antecedent-consecvent ce descriu o serie de implicaţii logice), axiome şi evenimente (modul de schimbare al atributelor sau al relaţiilor).

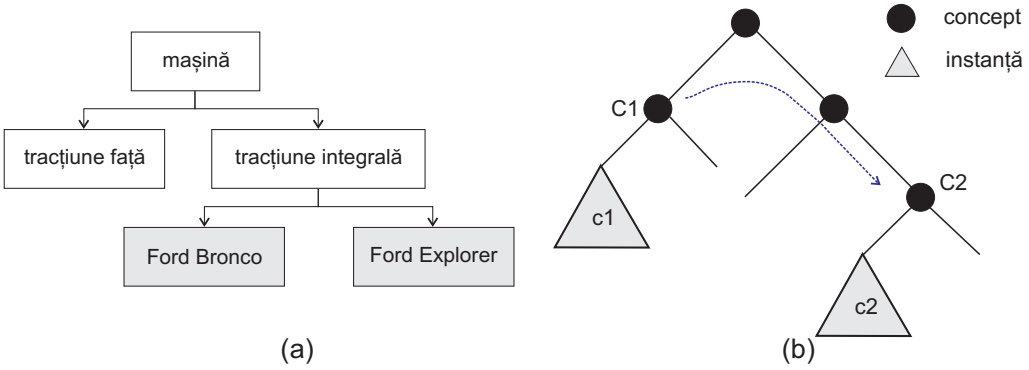


Figura 5.2: Exemple de ontologii: (a) definirea clasei "maşină" şi a obiectelor "Ford Bronco" şi "Ford Explorer" (exemplu din Wikipedia), (b) calculul distanţei dintre două concepte, C_1 şi C_2 (exemplu din [Worring 03]).

Un exemplu este ilustrat în Figura 5.2.(a) unde este prezentată o ontologie simplificată pentru clasa "maşină". O clasă poate fi subordonată unei alte clase, de exemplu clasa "maşină" poate fi considerată subclasă pentru clasa "autovehicul", deoarece toţi membrii acesteia sunt implicit şi membrii clasei "autovehicul", sau la rândul ei poate să conţină alte clase subordonate,

de exemplu clasele "tracţiune faţă" şi "tracţiune integrală" în exemplul din figură.

Acest mod de reprezentare crează o structură ierarhică în care la nivelul ierarhic superior se găsesc clasele cele mai generale iar la cel inferior clasele cele mai specifice. Relaţiile de subordonare implică mostenirea proprietăţilor claselor superioare ("părinţi") către clasele inferioare ("copii"). O partiţie a ontologiei reprezintă un set de clase şi regulile asociate acestora ce asigură faptul că obiectele pot fi clasificate în subclasa cea mai apropiată.

De exemplu, Figura 5.2.(a) conţine de fapt diagrama parţială a unei ontologii ce corespunde unei partiţii a clasei "maşină" în clasele "tracţiune faţă" şi "tracţiune integrală". Regula de partiţionare determină dacă o anumită maşină poate fi clasificată în una dintre cele două subclase. În acest mod de reprezentare, obiectele sunt descrise de atribute. Tipul unui obiect şi tipul atributelor determină modul de relaţionare între acestea. O relaţie dintre un obiect şi un atribut reflectă faptul că acesta este specific obiectului de care este relaţionat.

În exemplul din Figura 5.2.(a), obiectul "Ford Explorer" poate conţine atribute de tipul:

- <se numeşte> Ford Explorer,
- <are drept componentă> uşă (număr minim şi maxim 4),
- <are drept componentă unul dintre> {motor 4.0 litrii, motor 4.6 litrii},
- <are drept componentă> transmisie cu 6-viteze,

Mai multe informaţii relative la ontologii pot fi găsite în [Gómez-Pérez 04].

În cazul comparării descrierilor semantice reprezentate sub formă de concepte, o metodă simplă constă în evaluarea distanţei ce trebuie parcursă în arborele unei ontologii pentru a ajunge de la un concept la altul. Un exemplu este prezentat în Figura 5.2.(b) în care am ilustrat conceptul C_1 şi C_2 în contextul unei ontologii. Având la dispoziţie instanţele c_1 şi respectiv c_2 ale acestor concepte, obţinute de exemplu din datele multimedia ce trebuiesc comparate, o măsură a similarităţii acestora poate fi determinată ca numărul de paşi necesari în arbore pentru a ajunge de la conceptul C_1 la C_2, şi anume 3 pentru acest exemplu (ce corespunde numărului de laturi ale arborelui ce trebuiesc parcurse, vezi săgeată în figură).

CAPITOLUL 6

Tehnicile de tip "relevance feedback"

Aşa cum a fost prezentat şi în Secţiunea 2.4, conceptul de "relevance feedback" în contextul sistemlor de indexare după conţinut se referă la interacţia cu utilizatorul în vederea îmbunătăţirii rezultatelor iniţiale ale căutării. În general, mecansimul de "relevance feedback" funcţionează după următorul algoritm [Manning 08]:

1. **căutarea datelor dorite**: utilizatorul realizează o anumită căutare specificând datele dorite prin formularea unei cereri de căutare ("query"). Sistemul, pe baza mecanismului implementat, returnează rezultatele ce au caracteristicile cele mai apropiate de "query" folosind un anumit criteriu de similaritate. Până în acest punct, procesul este identic procesului de căutare al unui sistem de indexare (vezi Secţiunea 2);

2. **evaluare rezultate** de către utilizator: în funcţie de performanţa sistemului, rezultatele obţinute pot fi mai mult sau mai puţin relevante pentru datele căutate. În acest punct, utilizatorul analizează rezultatele şi le clasifică manual ca fiind, fie relevante pentru căutare (rezultat corect), fie ne-relevante (rezultat eronat). De regulă acest proces are loc pentru un număr limitat de rezultate, de ordinul zecilor;

3. **rafinare rezultate**: informaţiile furnizate de utilizator sunt folosite drept referinţă ("ground truth"). Pe baza acestora, sistemul va recalcula o reprezentare mai bună a rezultatelor căutării furnizând utilizatorului o rafinare a rezultatelor în funcţie de asemănarea cu datele

indicate drept relevante pentru căutare. Acest pas constituie practic algoritmul efectiv de "relevance feeback";

4. **re-iterare algoritm**: în funcţie de calitatea noilor rezultate obţinute, întreg procesul poate fi repretat prin reluarea punctului 2 până când rezultatele obţinute sunt satisfăcătoare pentru utilizator sau îndeplinesc un anumit criteriu de performanţă.

Un exemplu este prezentat în Figura 6.1 pentru căutarea de imagini în baza de date Microsoft Object Class Recognition[1].

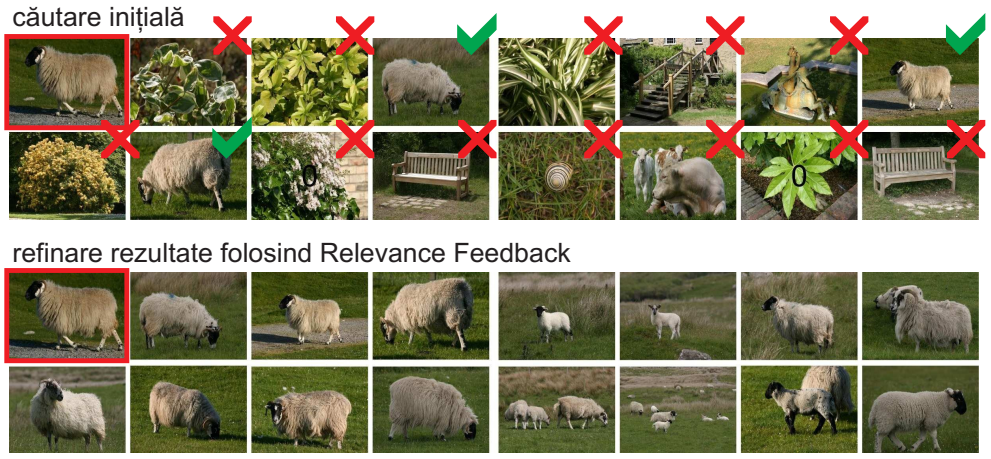


Figura 6.1: Exemplu de "relevance feedback" în cazul căutării de imagini (metoda propusă în [Mironică 12a]). Imaginile de mai sus reprezintă rezultatele sistemului de indexare pentru căutarea imaginilor similare cu imaginea marcată de chenarul roşu ($\surd$ şi x reprezintă rezultatele corecte şi respectiv eronate marcate de utilizator) în timp ce imaginile din partea de jos reprezintă rafinarea rezultatelor cu "relevance feedback".

Cererea de căutare a fost formulată prin furnizarea unei imagini exemplu (vezi imagine marcată de dreptunghiul roşu). În prima parte a figurii sunt ilustrate rezultatele obţinute iniţial de sistemul de indexare (din motive de spaţiu sunt prezentate doar primele 15 rezultate; rezultatele sunt afişate în ordine descrescătoare a similarităţii, de la stânga la dreapta şi de sus în jos). Pentru aceste rezultate utilizatorul a marcat imaginile corecte (simbol $\surd$) şi respectiv cele eronate (simbol x).

În partea de jos a figurii sunt prezentate rezultatele rafinate în urma aplicării metodei de "relevance feedback" ierarhic propusă în [Mironică 12a].

[1]vezi notă de subsol 1.

Se poate observa o îmbunatăţire semnificativă a performanţelor sistemului, imaginile returnate în acest caz corespunzând în totalitate cererii de căutare.

În funcţie de modul în care sunt preluate informaţiile de la utilizator ("feedback"), întâlnim trei tipuri de algoritmi:

- **"feedback" explicit**: corespunde algoritmului descris anterior în care utilizatorul el însuşi specifică care dintre rezultate sunt corecte şi care sunt eronate;
- **"feedback blind"** sau pseudo-feedback: presupune simularea interacţiei cu utilizatorul şi se bazează pe ipoteza conform căreia sistemul de indexare este suficient de performant (prin prisma descriptorilor de conţinut folosiţi şi al mecanismului de căutare) astfel încât este foarte probabil ca primele rezultate returnate să conţină un număr semnificativ de rezultate corecte. În acest caz, interacţia cu utilizatorul este substituită prin considerarea implicită a primelor k rezultate drept relevante [Larson 10]. Pe măsură ce datele căutate devin din ce în ce mai complexe (exemplu multimodale), ipoteza de relevanţă a primelor rezultate devine din ce în ce mai dificil realizabilă ceea ce conduce la performanţe limitate pentru această abordare;
- **"feedback" indirect**: interacţia cu utilizatorul se realizează în acest caz în mod indirect, pe baza observării "comportamentului" de căutare a diverşi utilizatori în situaţii diferite. De exemplu, sistemul poate utiliza informaţii despre datele pe care utilizatori diferiţi le-au accesat în urma căutării unor documente cu conţinut asemănător (faptul că documentele respective au fost accesate conferă un grad de încredere ridicat privind relevanţa conţinutului acestora) [Kelly 03]. Aceste informaţii pot fi stocate cu uşurinţă de motoarele de căutare actuale şi în special de cele "on-line" bazate pe text, ca de exemplu istoricul căutării pe Internet ce implică accesarea de documente web, mesagerie electronică, articole de ştiri, filme, cărţi, programe TV şi aşa mai departe.

În funcţie de durata relativă a procesului de antrenare a sistemului, algoritmii de "relevance feedback" se împart în algoritmi cu *antrenare cu termen scurt de învăţare* ("short-term relevance feedback") şi respectiv cu *antrenare cu termen lung* ("long-term relevance feedback").

Antrenarea cu termen scurt de învăţare presupune interacţia cu utilizatorul doar în sesiunea curentă fiind şi categoria cea mai studiată de metode în contextul sistemelor de indexare multimedia. Dintre abordările cele mai frecvent folosite putem enumera: algoritmi de schimbare a punctului de interogare, algoritmi de determinare a importanţei descriptorilor de conţinut,

algoritmi statistici sau algoritmi ce implementează procesul de "relevance feedback" ca o problemă de clasificare binară a datelor în două clase de relevanţă: date relevante şi date ne-relevante pentru utilizator. O parte dintre aceste metode sunt detaliate ulterior în secţiunile următoare.

Principalele provocări ale acestui tip de abordare pot fi sintetizate cu următoarele:

- numărul rezultatelor căutării pentru care utilizatorul furnizează relevanţa acestora este de regulă mult mai mic decât dimensiunea descriptorilor de conţinut (dimensiunea spaţiului de caracteristici) folosiţi pentru reprezentarea datelor (vezi Secţiunea 2.1), oferind astfel o capacitate de selecţie limitată din punct de vedere statistic;

- realizarea interacţiei cu utilizatori diferiţi va conduce în general la rezultate diferite şi uneori chiar contradictorii. Persoane diferite au moduri de percepţie diferită cu privire la proprietăţile aceloraşi concepte, de exemplu un expert va percepe într-un mod diferit conţinutului imaginii unei opere de artă faţă de o persoană neavizată. Acest lucru va conduce la varierea performanţelor sistemului de "relevance feedback" în funcţie de utilizator;

- discrepanţa dintre numărul de rezultate relevante şi cele nerelevante. De cele mai multe ori numărul de rezultate relevante returnate de sistem tinde să fie foarte mic nefiind suficiente pentru ca sistemul să se poată adapta la acestea. Acceaşi problemă apare şi în situaţia opusă, când nu există practic nici un rezultat nerelevant, situaţie în care sistemul nu poate face diferenţa dintre cele două cazuri;

- rafinarea rezultatelor în timp real este de asemenea un punct critic. Având în vedere interacţia directă cu utilizatorul, pentru a fi rentabil, sistemul trebuie să poată furniza noile rezultate cât mai rapid, implicând un timp de aşteptare minim din partea utilizatorului.

Învăţarea de lungă durată se foloseşte nu numai de informaţiile obţinute de la utilizator în sesiunea curentă, ci de toate informaţiile furnizate de-a lungul timpului de utilizatori diferiţi şi în sesiuni diferite. Acestea sunt de regulă stocate de cele mai multe ori sub forma unei reprezentări matriceale a relaţiilor descoperite ca existând între informaţiile din baza de date, relaţii ce sunt actualizate pe măsură ce se obţin noi informaţii de la utilizatori.

Ca şi în cazul anterior, există o serie de limitări ale acestui mod de abordare, cele mai semnificative fiind:

- aceşti algoritmi sunt mai dificil de implementat în cazul bazelor de date ce presupun frecvent eliminarea şi adăugarea de date noi;

- gradul de succes depinde foarte mult de cantitatea de informaţii de "feedback" stocate anterior, de cele mai multe ori în realitate preferându-se utilizarea unei combinaţii între strategii de învăţare de scurtă şi lungă durată;

- datorită utilizării mai multor surse de "feedback" informaţia stocată tinde să fie neomogenă şi foarte probabil să nu acopere toate datele;

- ca şi în cazul anterior, procesul de rafinare trebuie să poată fi implementat în timp real. Suplimentar complexităţii datelor de prelucrat, sistemul trebuie să fie capabil să analizeze şi un volum semnificativ de date de "feedback" de la utilizatori. De regulă pentru a soluţiona această problemă, se preferă împărţirea bazei de date pe diverse niveluri de relevanţă folosind ierarhii arborescente de conţinut.

6.1 Algoritmul Rocchio

Algoritmii de schimbare a punctului de interogare constituie una dintre primele abordări de tip "relevance feedback" ale problemei rafinării rezultatelor căutării, dezvoltate iniţial în contextul căutării de documente textuale, exemplu fiind algoritmul propus de Rocchio [Rocchio 71]. Pornind de la modul de reprezentare al datelor într-un sistem clasic de indexare după conţinut în care fiecare document este reprezentat ca un punct în spaţiul de caracteristici definit de descriptorii de conţinut asociaţi (vezi şi Figura 2.2), o anumită cerere de căutare a utilizatorului ("query") este descrisă la rândul ei în acelaşi spaţiu sub forma unui punct numit şi punct de interogare.

Acest lucru este ilustrat schematic în Figura 6.2. Axele a_1, a_2, ..., a_n reprezintă valorile atributelor de conţinut ce definesc spaţiul de caracteristici n-dimensional. Fiecare punct reprezintă valorile descriptorilor unui document din baza de date. Cererea de căutare este reprezentată în acest caz de dreptunghiul verde (punctul de interogare). În urma procesului de căutare, sistemul returnează ca rezultat datele cele mai apropiate punctului de interogare marcate în Figura 6.2 de cercul punctat (punctele care se află la o anumită distanţă de "query", de regulă în interiorul unei sfere). Aceste rezultate sunt prezentate utilizatorului de regulă în ordinea descrescătoare a distanţei faţă de punctul de interogare.

Conform algoritmului de "relevance feedback", utilizatorul marchează mai departe rezultatele ca fiind, fie relevante, fie nerelevante pentru datele căutate; de exemplu punctele marcate în figură cu cercuri verzi şi respectiv punctele marcate cu "+" de culoare roşie.

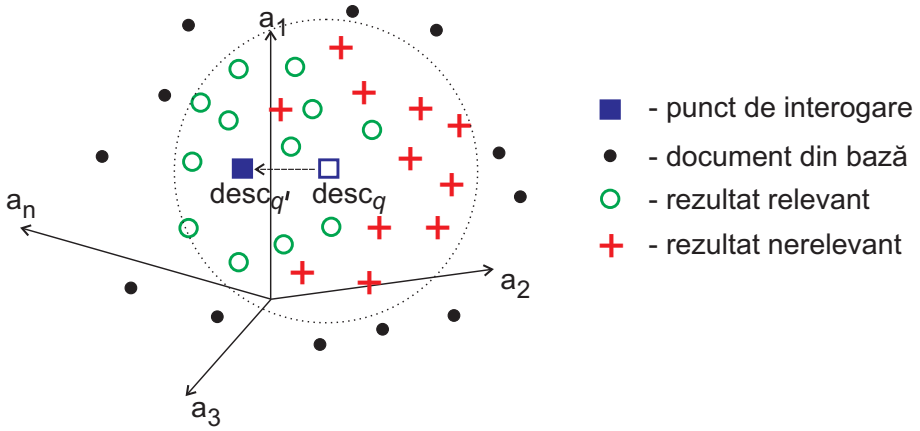


Figura 6.2: Modul de schimbare a punctului de interogare în cazul metodei propuse de Rocchio [Rocchio 71] (reprezentarea obiectelor din bază în spaţiul de caracteristici; $desc_q$ reprezintă punctul de interograre iniţial iar $desc_{q'}$ noul punct de interogare calculat - notaţiile sunt explicate în text).

Algoritmul lui Rocchio utilizează mulţimea de documente relevante, R, şi respectiv de documente nerelevante, N, pentru a redefini un nou punct de interogare folosind următoarea relaţie:

$$desc_{q'} = \alpha \cdot desc_q + \beta \cdot \frac{1}{||R||} \sum_{desc_i \in R} desc_i - \gamma \cdot \frac{1}{||N||} \sum_{desc_j \in N} desc_j \qquad (6.1)$$

unde $desc_{q'}$ reprezintă noul punct de interogare, $desc_q$ reprezintă punctul de interogare iniţial, α (ponderea punctului iniţial de interogare), β (factorul de importanţă al rezultatelor relevante) şi γ (factorul de importanţă al rezultatelor nerelevante) sunt o serie de ponderi alese empiric (valorile acestora sunt cuprinse în intervalul $[0; 1]$), $||.||$ este operatorul ce returnează numărul de elemente ale unei mulţimi iar $desc = \{a_1, ..., a_n\}$ reprezintă descriptorii de conţinut ai rezultatelor căutării.

Definit în acest fel, noul punct de interogare tinde să se deplaseze spre centroidul mulţimii R a rezultatelor marcate ca fiind relevante, ceea ce în urma reluării mecanismului de căutare va conduce la rezultate mai relevante.

6.2 Estimarea importanţei atributelor

Algoritmii de estimare a importanţei atributelor ("Feature Relevance Estimation") [Rui 99] pleacă de la ipoteza conform căreia pentru o anumită căutare

ponderea atributelor descriptorilor de conţinut poate influenţa relevanţa rezultatelor. În mod implicit, atributele descriptorilor sunt considerate a avea o contribuţie identică la localizarea datelor celor mai similare, acest lucru fiind realizat pe baza calculului unei măsuri de distanţă (de exemplu distanţa Euclidiană, vezi şi Secţiunea 5). Pe baza interacţiei cu utilizatorul, ponderile atributelor pot fi modificate astfel încât să se îmbunătăţească rezultatele căutării.

Folosind notaţiile anterioare, dacă $desc = \{a_1, ..., a_n\}$ reprezintă descriptorul de conţinut al datelor, unde a_i cu $i = 1, ..., n$ reprezintă valorile atributelor acestuia, atunci se va considera un anumit vector de ponderi, $W = \{w_1, ..., w_n\}$, unde w_i reprezintă ponderea atributului a_i. Aceste valori sunt iniţial considerate egale cu 1 (cu alte cuvinte nu există ponderare).

Sistemul de indexare realizează căutarea datelor pe baza comparării descriptorilor şi returnează rezultatele în ordinea descrescătoare a similarităţii. Ca şi în cazul anterior, utilizatorul marchează rezultatele relevante şi respectiv nerelevante. Pe baza acestor informaţii se va modifica ponderea individuală a fiecărui atribut.

O variantă o reprezintă calculul lui w_i în funcţie de abaterea pătratică medie a valorilor atributelor σ_i, şi anume:

$$w_i = \frac{1}{\sigma_i} \tag{6.2}$$

unde σ_i reprezintă abaterea pătratică medie a valorilor atributului a_i pentru documentele marcate drept relevante de utilizator. Definit în acest fel, un atribut cu grad de importanţă ridicat va tinde să aibă o valoare relativ constantă pentru fiecare document în timp ce un atribut mai puţin discriminant pentru datele căutate va tinde să aiba o gamă mult mai mare de valori, ponderea acestuia fiind redusă proporţional.

O altă abordare constă în folosirea de ponderi ce depind de rezultatele căutării individuale după fiecare atribut în parte:

$$w_i = \frac{2 \cdot ||R_i||}{T} \tag{6.3}$$

unde R_i reprezintă mulţimea documentelor relevante în cazul unei căutări folosind drept descriptor doar atributul a_i, $||.||$ este operatorul ce returnează numărul de elemente ale unei mulţimi iar T reprezintă numărul total de documente relevante din bază.

Odată determinate ponderile atributelor, acestea sunt folosite la rafinarea rezultatelor căutării prin calcularea similarităţii documentelor pe baza unei

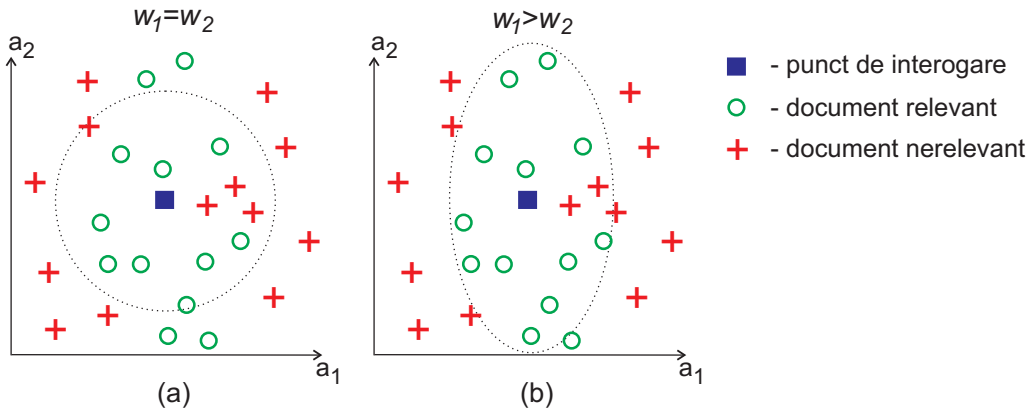


Figura 6.3: Estimarea importanţei atributelor cu metoda [Rui 99] (reprezentarea obiectelor din bază în spaţiul de caracteristici - pentru exemplificare s-au ales doar două atribute): (a) reprezentarea rezultatelor căutării (delimitate de cercul punctat), (b) modificarea rezultatelor în funcţie de noua pondere a atributelor (delimitate de elipsa punctată).

măsuri de distanţe ponderate:

$$d_{FRE}(desc_x, desc_y, W) = \sqrt{\frac{\sum_{i=1}^{n} w_i \cdot (a_{xi} - a_{yi})^2}{\sum_{i=1}^{n} w_i}} \tag{6.4}$$

unde $desc_x$ şi $desc_y$ reprezintă descriptorii de conţinut a două documente iar a_{xi} şi a_{yi} cu $i = 1, ..., n$ atributele acestora.

Modificarea ponderilor asociate fiecărui atribut individual al descriptorului în funcţie de rezultatele relevante se traduce în spaţiul de caracteristici prin modificarea regiunii de selecţie a rezultatelor de la o sferă la un elipsoid, adaptându-se mulţimii de documente relevante. Acest lucru este ilustrat schematic în Figura 6.3.

6.3 Support Vector Machines

Motivaţi de succesul implementării tehnicilor de învăţare asistată de calculator ("machine learning") în contextul sistemelor de indexare după conţinut, algoritmii de clasificare şi-au găsit aplicabilitate şi în cazul tehnicilor de "relevance feedback". Astfel, problema îmbunătăţirii performanţelor sistemului de căutare pe baza utilizării informaţiei furnizate de utilizator este transformată într-o problemă clasică de clasificare.

Documentele marcate ca fiind relevante şi respectiv nerelevante sunt folosite pentru antrenarea unui anumit clasificator care să permită catalogarea datelor în una dintre cele două clase: documente relevante şi respectiv documente nerelevante. Mai departe, documentele din bază sunt trecute prin clasificator şi vor fi astfel re-alocate uneia dintre cele două clase. Practic, informaţia de la utilizator este folosită pe post de "ground truth"[2] pentru determinarea unui set de reguli care să permită partiţionarea datelor în cele două clase de relevanţă. În urma clasificării, datele vor primi un nou rang calculat în funcţie de un grad de relevanţă atribuit de clasificator, ceea ce conduce global la rafinarea rezultatelor iniţiale.

Dintre tehnicile de clasificare a datelor cel mai frecvent întâlnite în contextul de "relevance feedback" putem menţiona: Support Vector Machines (SVM), k-Nearest Neighbors (kNN) sau arborii de decizie (ca de exemplu Random Forests). Pentru mai multe detalii relativ la tehnicile de clasificare a datelor cititorul se poate raporta la [Ionescu 09] [Witten 05] (vezi şi explicaţia din Secţiunea 4.2).

În cele ce urmează ne vom limita la prezentarea unuia dintre algoritmii de clasificare foarte populari care s-a dovedit eficient în rezolvarea diferitelor probleme de indexare a conţinutului multimedia şi anume Support Vector Machines.

Support Vector Machines (SVM) realizează clasificarea datelor prin construcţia unui hiperplan[3] ce separă în mod optimal datele de intrare în două categorii [Welling 05]. Aceasta este o problemă de clasificare liniară. Având în vedere că există o multitudine de hiperplane ce pot separa datele, SVM restricţionează căutarea la acele hiperplane ce permit o separare maximă între cele două clase (maximizarea "marginii" dintre date).

Cu alte cuvinte, se caută hiperplanul cu proprietatea ca acesta să maximizeaze distanţa faţă de cel mai apropiat punct din spaţiul de caracteristici. Acesta este denumit "hiperplanul marginii maximale" ("maximum-margin hyperplane"). Un exemplu este prezentat în Figura 6.4.(a) unde spaţiul de caracteristici este separat de hiperplanul H_1 ce nu permite separarea datelor, hiperplanul H_2 care are o margine redusă şi respectiv hiperplanul H_3 ce maximizează separarea dintre clase (vezi distanţele faţă de cele mai apropiate puncte).

Formalizarea problemei de clasificare abordată de SVM este următoarea:

[2] vezi notă de subsol 9.

[3] un hiperplan este un concept folosit în domeniul algebrei liniare pentru a generaliza noţiunea de linie - folosită în geometria Euclidiană a planului, sau de plan - folosită în geometria Euclidiană tridimensională, pentru cazul n-dimensional, cu $n > 3$.

având la dispoziţie un set de date de antrenare, D, constituit ca fiind:

$$D = \{(X_i, c_i) | X_i \in R^n, c_i \in \{-1, 1\}\} \tag{6.5}$$

unde X_i este un vector $n-$dimensional (de exemplu descriptorul datelor), c_i indică clasa din care face parte vectorul X_i (valori etichete -1 şi 1), i reprezintă indicele vectorului curent, cu $i = 1, ..., p$, iar p reprezintă numărul de vectori consideraţi; se caută hiperplanul marginii maximale ce permite separarea punctelor din clasa $c_i = 1$ de cele din clasa $c_i = -1$ (vezi Figura 6.4.(b)).

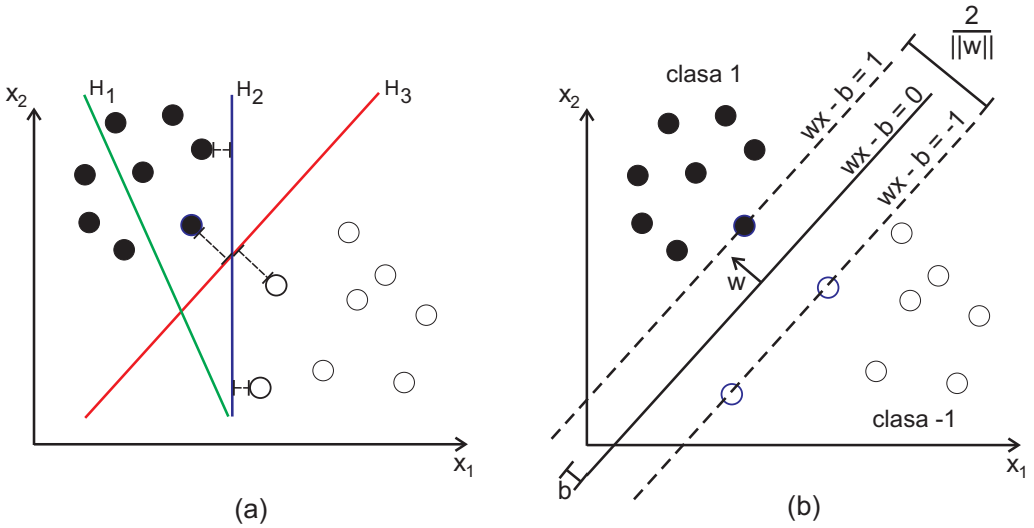


Figura 6.4: Principiul SVM (cercurile reprezintă vectorii de caracteristici, X_1 şi X_2 formează spaţiul de caracteristici): (a) exemple de hiperplane de separare a datelor, (b) hiperplanul marginii maximale în cazul a două clase (sursă exemple Wikipedia).

Un hiperplan oarecare poate fi definit ca un set de puncte X ce satisfac următoarea relaţie:

$$W \cdot X - b = 0 \tag{6.6}$$

unde W reprezintă un vector normal (perpendicular pe hiperplan), $\cdot$ reprezintă produsul scalar iar parametrul $\frac{b}{||W||}$ va defini decalajul hiperplanului faţă de originea axei de coordonate, de-a lungul vectorului W (vezi Figura 6.4.(b)).

În scopul definirii marginii maximale, căutăm valorile lui W şi b astfel încât acestea să maximizeze distanţa dintre hiperplanele paralele, cele mai

depărtate, dar care încă separă datele. Acestea sunt date de ecuaţiile:

$$W \cdot X - b = 1 \tag{6.7}$$
$$W \cdot X - b = -1 \tag{6.8}$$

Distanţa dintre acestea este $\frac{2}{||W||}$, astfel că problema maximizării se transformă într-o problemă de minimizare a valorii $||W||$. De asemenea, pentru a preveni ca punctele să se găsească pe margini, se folosesc o serie de constrângeri suplimentare, astfel marginea maximală este determinată de condiţiile următoare:

$$W \cdot X_i - b \geq 1, \quad X_i \in c_1 \tag{6.9}$$
$$W \cdot X_i - b \leq -1, \quad X_i \in c_{-1} \tag{6.10}$$

sau

$$c_i \cdot (W \cdot X_i - b) \geq 1 \tag{6.11}$$

pentru oricare $i \in [1; p]$.

Transformată într-o problemă de optimizare, clasificarea SVM poate fi enunţată astfel: alege parametrii W şi b astfel încât să minimizeze valoarea $||W||$ cu constrângerea ca: $c_i \cdot (W \cdot X_i - b) \geq 1$, pentru oricare i. Această clasificare este valabilă totuşi doar în cazul în care datele sunt liniar separabile. În realitate, mai ales în contextul descriptorilor de conţinut multimodali, este puţin probabil ca separarea acestora să se poată realiza liniar.

Pentru a crea un clasificator SVM neliniar, la maximizarea marginii dintre clase se folosesc ceea ce numim funcţii nucleu sau "kernel functions". Operaţiile de înmulţire scalară sunt înlocuite acum de nuclee de funcţii neliniare, $k(X, X')$ unde X şi X' sunt doi vectori. În acest fel, hiperplanul marginii maximale va fi potrivit datelor într-un spaţiu de caracteristici transformat neliniar. Dintre nucleele cel mai frecvent folosite putem menţiona:

- nucleu polinomial omogen:

$$k(X, X') = (X \cdot X')^d \tag{6.12}$$

  unde d este un număr întreg;

- nucleu polinomial neomogen:

$$k(X, X') = (X \cdot X' + 1)^d \tag{6.13}$$

- funcţie radială:

$$k(X, X') = \exp(-\gamma ||X - X'||^2) \tag{6.14}$$

  unde $\gamma > 0$;

- funcţie radială Gaussiană:

$$k(X, X') = \exp\left(-\frac{||X - X'||^2}{2\sigma^2}\right) \tag{6.15}$$

  unde σ^2 reprezintă varianţa statistică;

- funcţie sigmoidă:

$$k(X, X') = \tanh(\kappa \cdot X \cdot X' + c) \tag{6.16}$$

  unde $tanh(.)$ reprezintă tangenta hiperbolică, $\kappa > 0$ iar $c < 0$.

Cu toate că SVM este un clasificator binar, acesta poate fi folosit cu succes pentru a rezolva probleme de clasificare multi-clasă specifice indexării după conţinut. Una dintre metodele cele mai uzuale constă în transformarea clasificării multi-clasă într-o succesiune de clasificări binare [Kotsiantis 07] (de exemplu folosind clasificatori binari ce clasifică o clasă faţă de toate celelalte - "one-versus-all", sau care clasifică fiecare pereche de clase - "one-versus-one").

În cele ce urmează vom prezenta un studiu comparativ al performanţelor a o serie de algoritmi de "relevance feedback" în contextul unei căutări de imagini folosind descriptori de conţinut. Algoritmii vizaţi sunt: Rocchio, Relevance Feature Estimation, Support Vector Machines (SVM), arbori de decizie (TREE), AdaBoost (BOOST), Random Forests şi clasificare ierarhică [Mironică 12b]. Testele sunt efectuate folosind o bază "off-line" şi anume Microsoft Object Class Recognition[4] ce conţine imagini cu 23 de categorii de obiecte (de exemplu animale, persoane, avioane, maşini şi aşa mai departe). Căutarea presupune identificarea tuturor imaginilor ce conţin un anumit obiect furnizat de utilizator.

Rezultatele sunt prezentate în Figura 6.5. Graficele ilustrează performanţa căutării pe baza valorii MAP (Mean Average Precision, vezi Secţiune 8; reprezentată pe axa oY - valoarea maximă este 100 ce indică o performanţă de 100%) raportată la numărul de sesiuni de "feedback" ale utilizatorului (axa oX; vezi explicaţia de la începutul Secţiunii 6). Rezultatele prezentate sunt obţinute folosind descriptori de culoare clasici (vezi Secţiune 3.1).

Ceea ce se observă imediat este faptul că performanţa căutării creşte semnificativ cu numărul sesiunilor de "feedback", de exemplu cu până la 20% în cazul metodei de clasificare ierarhică (comparat cu rezultatele din prima sesiune). De asemenea, comparativ cu rezultatele obţinute fără a aplica

[4]vezi notă de subsol 1.

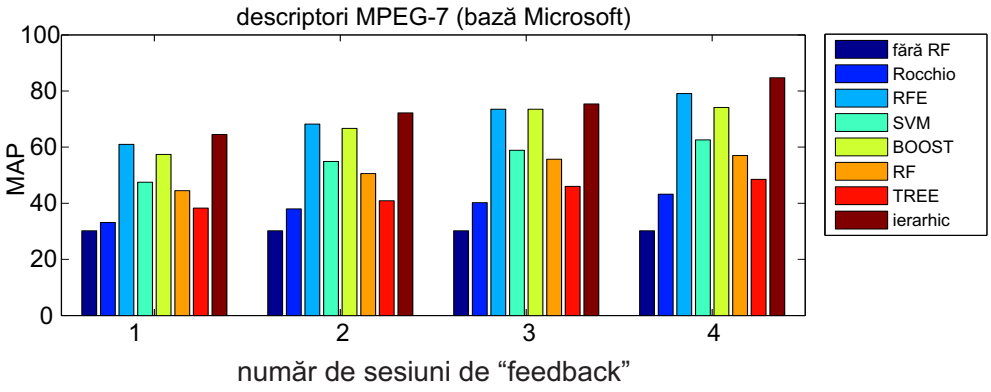


Figura 6.5: Compararea performanţelor tehnicilor de "relevance feedback" în contextul căutării de imagini [Mironică 12b] (notaţiile sunt explicate în text).

"relevance feedback", performanţa sistemului se poate dubla pentru o singură sesiune sau chiar ajunge la o performanţă de peste 80% după mai multe sesiuni, ceea ce este întradevăr un rezultat relevant. Creşterea semnificativă a performanţei se realizează de regulă pentru primele sesiuni de "feedback", în general după prima sesiune, urmând să se diminueze progresiv cu creşterea numărului de sesiuni. De exemplu, clasificarea ierarhică furnizează o creştere a performanţei cu 31% în prima sesiune (faţă de căutarea fără "relevance feedback") şi apoi doar de 7% (faţă de prima sesiune) şi de 4% faţă de a doua sesiune.

Din punct de vedere al metodelor, în ciuda superiorităţii clare a unor abordări faţă de altele, nu putem trage o concluzie generală, rezultatele fiind de regulă dependente de baza de test şi de sistemul de indexare folosit. În exemplul ilustrat, metoda de clasificare ierarhică urmată de estimarea importanţei atributelor (RFE) furnizează performanţele cele mai ridicate.

CAPITOLUL 7

Vizualizarea conţinutului multimedia

Vizualizarea conţinutului datelor multimedia este la rândul ei o problemă ce trebuie luată în calcul. Aceasta este integrată sistemului de navigare (vezi Sectiunea 2).

În contextul imaginilor, dificultatea vizualizării datelor ţine în cea mai mare parte de volumul de date ridicat ce trebuie accesat, conţinutul unei imagini putând fi reprezentat simplu prin reprezentarea acestuia la o rezoluţie scăzută (de exemplu pe bază de miniaturi). Astfel, o bază de imagini poate fi vizualizată eficient prin vizualizarea miniaturilor imaginilor din aceasta sub formă de planşe. În Figura 7.1 am prezentat ca exemplu modul de vizualizare folosit de platforma de căutare Flickr[1]. Se observă faptul că informaţia furnizată poate fi analizată foarte rapid de utilizator, timpul necesar fiind de ordinul zecilor de secunde.

În contextul secvenţelor de imagini, pe lângă volumul mare de date se mai adaugă şi problema vizualizării conţinutului video dinamic. Este evident faptul că vizualizarea în parte a fiecărei secvenţe este aproape imposibilă iar reprezentarea acestora cu o singură imagine este nerelevantă deoarce nu surprinde informaţia definitorie care ţine de conţinutul de mişcare şi de evoluţia în timp. O soluţie la această problemă constă în folosirea de rezumate de conţinut ce reprezintă practic modalităţi de reprezentare compactă a conţinutului, atât vizual cât şi temporal.

Tehnicile de rezumare automată a conţinutului video [Truong 07] vizează două categorii de rezumat, şi anume *rezumatul în imagini* (static) ce re-

[1] `http://www.flickr.com/search/?q=Tour+Eiffel&z=t`

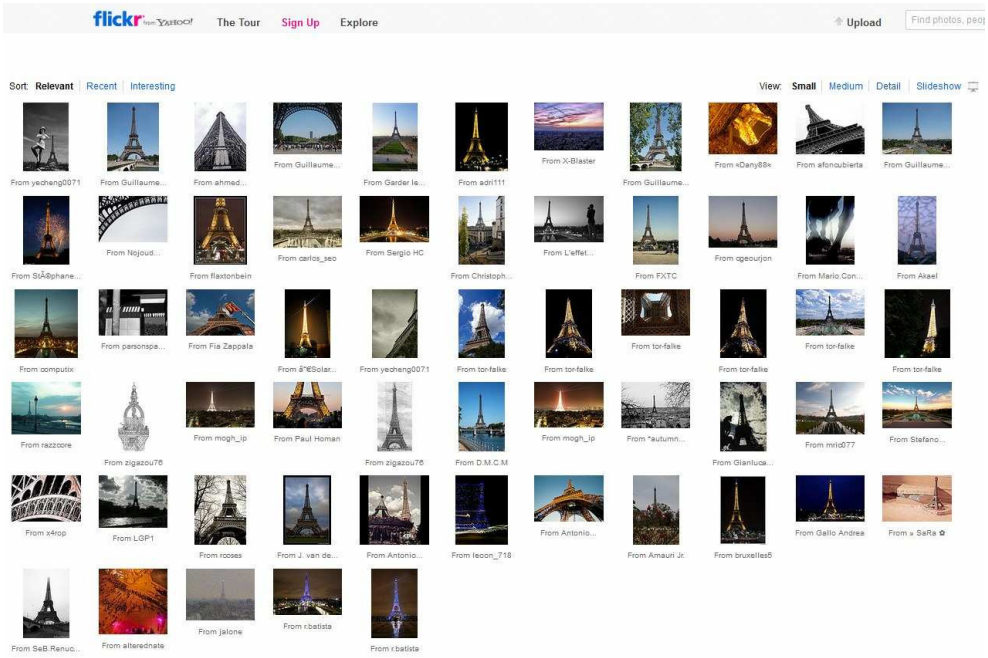


Figura 7.1: Exemplu de vizualizare a conţinutului unei colecţii de imagini pe platforma Flickr (rezultate obţinute în urma căutării de imagini cu "Tour Eiffel").

prezintă o colecţie de imagini reprezentative şi *rezumatul în mişcare* (dinamic), ce reprezintă o colecţie de pasaje reprezentative ale secvenţei. Rezumatele de conţinut permit utilizatorului să-şi facă rapid o idee globală asupra conţinutului secvenţei. Astfel, rezumatul static permite reprezentarea conţinutului vizual al secvenţei în doar câteva imagini (de exemplu câte o imagine pentru fiecare scenă reprezentativă), ce sunt uşor accesibile utilizatorului prin sistemul de navigare, timpul de vizualizare fiind neglijabil. Pe de altă parte, rezumatul dinamic aduce un plus de informaţie la nivelul acţiunii prezente în secvenţă, informaţie ce nu este disponibilă în rezumatul static. Totuşi, fiind el însuşi o secvenţă, în funcţie de nivelul de detaliu furnizat, timpul necesar vizualizării acestuia este mai ridicat decât în cazul rezumatului static, dar net inferior timpului de vizualizare integrală a secvenţei (un exemplu sunt rezumatele de tip "trailer" care tind să surprindă doar conţinutul de acţiune).

Pe lângă aspectul vizualizării propriu-zise a datelor, aşa cum am enunţat şi anterior, principala problemă a vizualizării colecţiilor multimedia este dată de necesitatea parcurgerii unui volum semnificativ de date, indiferent dacă

este vorba de imagini sau video. În cele ce urmează vom trece în revistă câteva sisteme de navigare multimedia ce întegrează tehnici inteligente de reprezentare a conţinutului datelor:

- MediaTable [Rooij 10] (vezi Figura 7.2): permite categorizarea imaginilor şi secvenţelor video. Sistemul foloseşte o vizualizare tabulară ce permite o vedere de ansamblu asupra colecţiei multimedia şi a descrierilor textuale ataşate cât şi o serie de interfeţe grafice ce permit sortarea, filtrarea, selectarea şi vizualizarea documentelor.

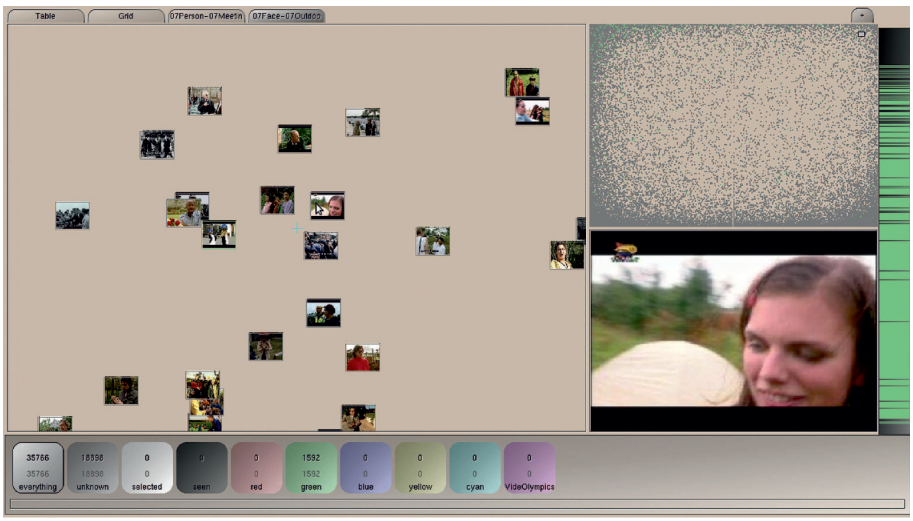


Figura 7.2: Sistemul MediaTable [Rooij 10].

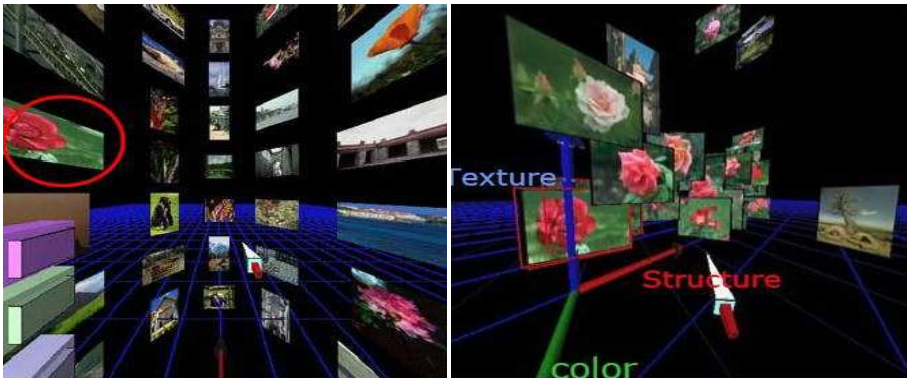


Figura 7.3: Sistemul 3D MARS [Nakazato 01].

În Figura 7.2 este ilustrat un exemplu de vizualizare a datelor după conţinut. Graficul din colţul din dreapta sus reprezintă o hartă a distribuţiei tuturor documentelor din bază în timp ce imaginea din colţul dreapta jos detaliază conţinutul documentului selectat curent.

- 3D MARS [Nakazato 01] (vezi Figura 7.3): permite vizualizarea colecţiilor de imagini folosind un sistem de reprezentare 3D de tip realitate virtuală. Imaginile sunt reprezentate în funcţie de conţinutul de culoare, textură şi respectiv structural (un exemplu este prezentat în imaginile din Figura 7.3).

- MediaMill Forkbrowser [Rooij 08] (vezi Figura 7.4): foloseşte un sistem de vizualizare intercalată atât a rezultatelor căutării video cât şi a conţinutului temporal. Pe axa de adâncime spre partea superioară sunt reprezentate rezultatele unei anumite căutări, pe axa orizontală este prezentat conţinutul temporal al unui segment al secvenţei curente ("timeline"), pe axele diagonale sunt ilustrate succesiuni de plane video al căror conţinut este similar cu imaginea vizualizată curent în centru ("similarity threads") iar pe axa de adâncime în partea de jos este prezentat istoricul căutărilor ("history").

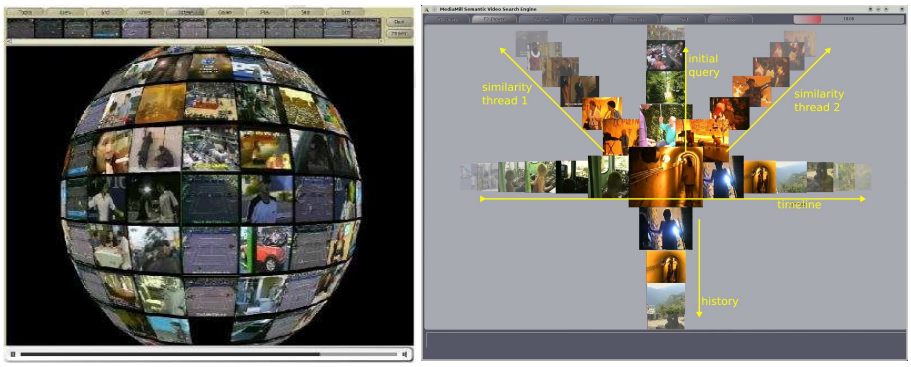


Figura 7.4: Sistemul MediaMill: Forkbrowser [Rooij 08].

- Reprezentare 3D cilindrică [Schoeffmann 11] (vezi Figura 7.5): permite reprezentarea colecţiilor de imagini sub forma unor reprezentări de tip "storyboard" ilustrate folosind o reprezentare cilindrică 3D. Diferite categorii de imagini sunt reprezentate folosind cilindrii diferiţi, utilizatorul putând selecta categoria dorită. Pentru vizualizarea curentă, imaginile prezentate în prim plan sunt reprezentate detaliat în timp ce imaginile din fundal sunt reprezentate schematic. Folosind interfaţa

grafică, utilizatorul poate derula imaginile rulate pe cilidru cât şi detalia o anumită regiune a cilindrului.

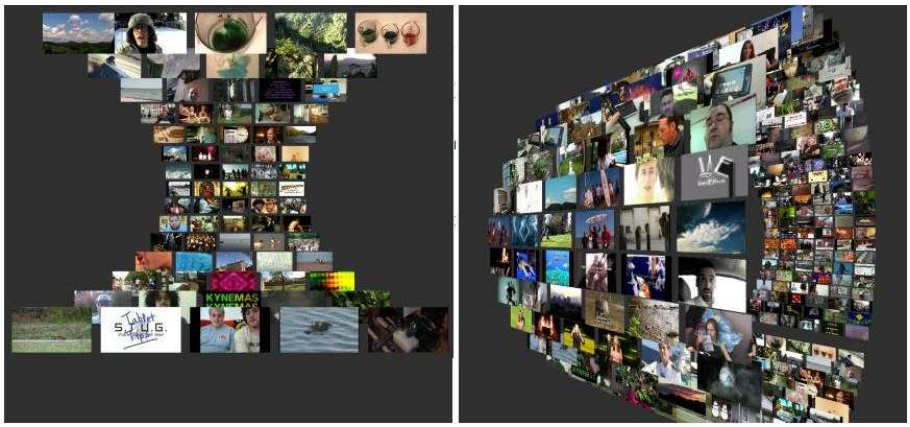


Figura 7.5: Sistem de reprezentare 3D cilindrică [Schoeffmann 11].

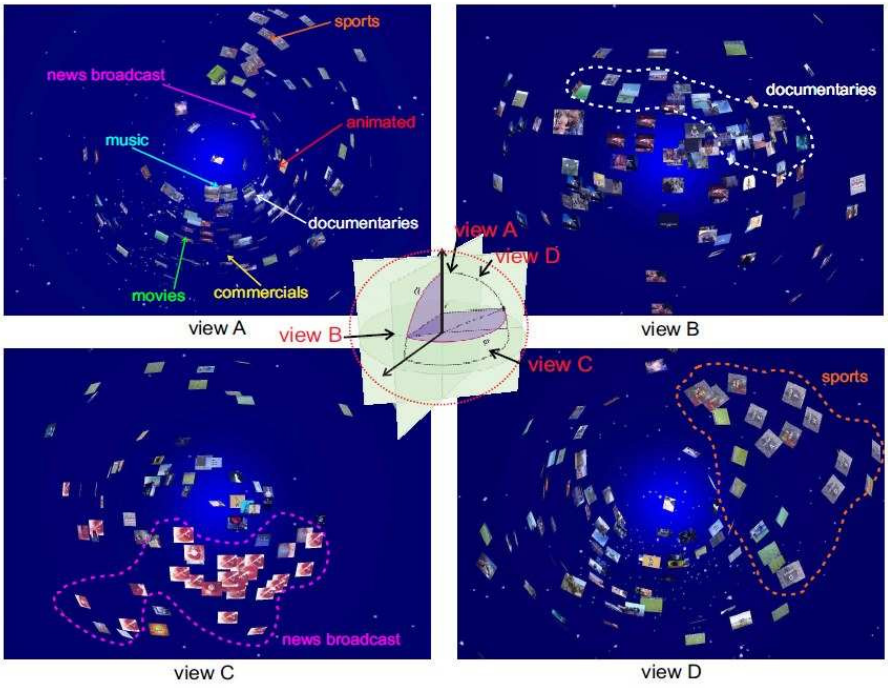


Figura 7.6: Sistemul MovieGlobe [Ionescu 12a].

Figura 7.7: Sistemul nepTunes [Knees 07].

- MovieGlobe [Ionescu 12a] (vezi Figura 7.6[2]): permite reprezentarea colecţiilor multimedia de imagini şi filme într-un spaţiu 3D virtual. Fiecare obiect multimedia este reprezentat ca un punct în acest spaţiu. Distribuţia obiectelor în spaţiul 3D este realizată în funcţie de similaritatea conţinutului acestora. Utilizatorul se poate deplasa virtual şi vizualiza conţinutul obiectelor întâlnite. În Figura 7.6 este prezentat un exemplu de reprezentare a filmelor în funcţie de gen (animaţie, sport, film, etc.).

- nepTune [Knees 07] (vezi Figura 7.7): permite vizualizarea conţinutului colecţiilor de muzică sub forma unor peisaje 3D virtuale pe care utilizatorul le poate explora. Peisajele "muzicale" sunt adaptate automat pe baza analizei conţinutului audio preferinţelor fiecărui utilizator.

[2]o demonstraţie este disponibilă la `http://imag.pub.ro/~bionescu/index_files/MovieGlobe.avi`

CAPITOLUL 8

Evaluarea perfomanţelor indexării

Aşa cum am menţionat şi în Capitolul 2.4, alături de problematica descrierii eficiente a conţinutului datelor cât şi a conceptului de similaritate între date de regulă heterogene, un aspect cel puţin la fel de important il constituie evaluarea performanţelor. Cu toate că un sistem de indexare poate funcţiona corect din punct de vedere al algoritmilor implementaţi şi al tehnicilor de reprezentare a datelor, acest lucru nu implică şi faptul că rezultatele obţinute sunt relevante pentru utilizator. Pentru validarea sistemului este necesară evaluarea globală a performanţelor acestuia, atât pentru seturi de date cât mai diverse cât şi pentru utilizatori diferiţi.

Metodele existente se împart în două categorii: metode de *evaluare subiectivă* ce au la bază utilizatorul şi respectiv metode de *evaluare obiectivă* ce se bazează pe calculul unor măsuri matematice. Acestea sunt descrise în cele ce urmează.

8.1 Evaluarea subiectivă

Campanii de evaluare. Evaluarea subiectivă a performanţelor implică însuşi utilizatorul. Practic calitatea rezultatelor obţinute de sistem este evaluată pe baza opiniei utilizatorilor (care până la urmă este chiar "consumatorul produsului"), ca de exemplu prin realizarea a ceea ce numim "user studies" (sau campanii de evaluare).

Utilizatorului i se pun la dispoziţie rezultatele obţinute de sistem şi acesta va completa un chestionar cu privire la gradul de satisfacţie şi relevanţa

acestora relativ la datele căutate. Procesul se repetă în general pentru un număr cât mai semnificativ de rezultate precum şi pentru cât mai mulţi utilizatori. De regulă, experimentele respectă un protocol bine definit şi sunt realizate în aceleaşi condiţii pentru toţi utilizatorii pentru a nu exista factori externi diferiţi care să influenţeze răspunsurile la întrebări. În final, răspunsurile obţinute relativ la performanţa sistemului sunt analizate din punct de vedere statistic şi se concluzionează asupra performanţelor medii globale ale sistemului.

Prezentăm pentru exemplificare o astfel de campanie de evaluare realizată în cazul tehnicilor de rezumare automată de conţinut. Sistemul evaluat este un sistem de generare automată a unui rezumat în imagini a unui document video [Ionescu 10] (o colecţie de imagini considerate ca fiind reprezentative pentru conţinutul secvenţei respective). Având în vedere subiectivitatea unui astfel de proces, se doreşte validarea acestuia de către utilizatori. Primul pas al campaniei constă în definirea protocolului de evaluare, şi anume acel algoritm pe care il vor urma utilizatorii. Definirea precisă a unui protocol asigură în primul rând standardizarea testului prin realizarea acestuia în acelaşi mod de către toţi participanţii la evaluare.

În cazul exemplului considerat protocolul folosit este unul simplu şi constă în următoarele etape: 1. vizualizarea într-o cameră de proiecţie a secvenţei video originale (izolarea utilizatorului de alte surse de informaţie si focalizarea asupra datelor evaluate), 2. prezentarea succesivă a imaginilor rezumatului propus (câte o imagine pe secundă), 3. completarea unui chestionar de către utilizator, 4. repetarea procesului pentru diverse secvenţe video. Chestionarul folosit cuprinde următoarele întrebări:

- întrebarea 1 - "În ce măsură estimaţi că rezumatul propus este relevant pentru conţinutul secvenţei?". Evaluarea acestei întrebări se realizează pe o scară de valori de la 0 la 10 cu următoarea semnificaţie: 0 nu ştiu, 1-2 deloc, 3-4 foarte puţin, 5-6 parţial, 7-8 în mare parte, 9-10 în totalitate. Pentru fiecare grad de apreciere sunt furnizate două niveluri;

- întrebarea 2 - "Cum estimaţi durata rezumatului din punct de vedere al numărului de imagini furnizate?". Evaluarea pentru această întrebare se realizează tot pe o scară de la 0 la 10 cu următoarea semnificaţie: 0 nu ştiu, 1-2 prea scurtă, 3-4 scurtă, 5-6 suficientă, 7-8 ridicată, 9-10 prea lungă.

În Figura 8.1 sunt prezentate rezultatele obţinute în urma testării rezumatelor pentru 10 secvenţe de animaţie (sursă [CITIA 13]) de către un număr de 27 de utilizatori. Graficele ilustrează scorul mediu obţinut pentru fiecare secvenţă şi întrebare în parte cât şi abaterea standard a acestor rezultate (un

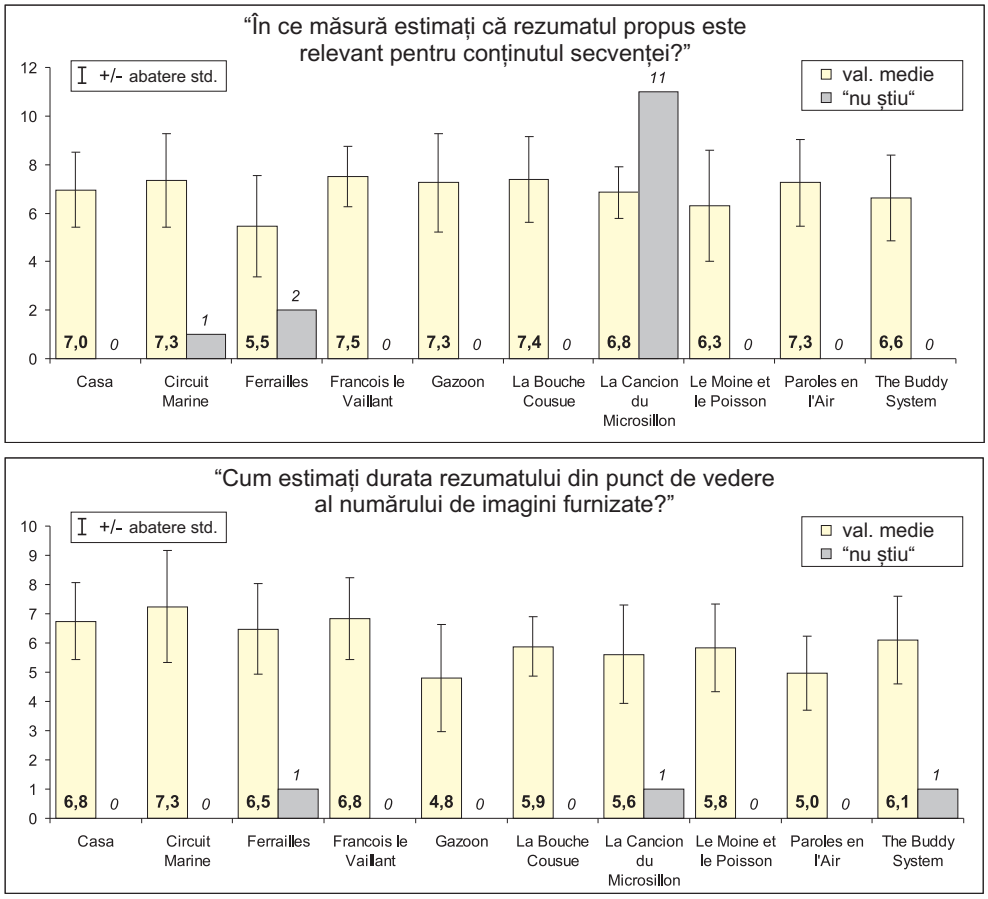


Figura 8.1: Exemplu de rezultate ale unei campanii de evaluare a performanţei în cazul metodei propuse în [Ionescu 10] (axa oX corespunde secvenţelor testate, axa oY corespunde scorului mediu furnizat de utilizatori, segmentele verticale ilustrează abaterea standard, bările gri reprezintă numărul de răspunsuri "nu ştiu" furnizate de utilizatori).

indicator al gradului de dispersie al răspunsurilor pentru utilizatori diferiţi şi implicit al subiectivităţii - cu cât această valoare este mai mare cu atât răspunsurile furnizate de utilizatori au fost mai diferite).

Ceea ce se observă imediat este faptul că rezultatele sunt dependente atât de utilizator cât şi de date. De exemplu, există situaţii în care utilizatorii nu pot furniza un răspuns relevant, de exemplu pentru secvenţa "La Cancion du Microsillon" numărul de răspunsuri "nu ştiu" este semnificativ (11 din 27); sau dispersia răspunsurilor este foarte ridicată ceea ce atestă un nivel ridicat de subiectivitate, de exemplu pentru secvenţa "Le Moine et le Poisson" unde

abaterea standard este de 2.3.

Totuşi, pe baza acestor date se poate concluziona la nivel global relativ la calitatea rezultatelor sistemului, în acest exemplu tehnica de rezumare propusă obţinând la întrebarea 1 un scor mediu global de 6.9, ceea ce corespunde faptului că este capabilă să reprezinte "în mare parte" conţinutul original; cât şi un scor mediu global de 6.1 la întrebarea 2 ceea ce corespunde faptului că durata rezumatului propus tinde să fie adecvată.

Crowd-sourcing. O alternativă actuală la realizarea fizică de campanii de evaluare o constituie folosirea mediului on-line şi anume a Internetului. Una dintre dificultăţile principale ale unei campanii de evaluare o constituie dificultatea de a dispune de un număr semnificativ de utilizatori la un anumit moment de timp într-o anumită locaţie. Astfel că o soluţie mai eficientă o constituie organizarea campaniei on-line, utilizatorii nefiind restricţionaţi a fi prezenţi fizic şi putând realiza evaluarea la momentul dorit în funcţie de disponibilitatea lor de timp. Mai mult, participarea on-line permite accesarea unui număr semnificativ de utilizatori din toată lumea.

Un domeniu aparte îşi găseşte în prezent aplicaţie în contextul sistemelor de evaluare a performanţelor algoritmilor multimedia şi anume acela de "crowd-sourcing". Cu toate că dezvoltarea "crowd-sourcing" nu este legată de acest context, fiind dezvoltată în principal pentru realizarea unei structuri de prestare de servicii la distanţă - conceptul de "crowd-sourcing" fiind definit ca "procesul de formulare a unei anumite sarcini de lucru, divizarea acesteia în micro-sarcini ce pot fi realizate foarte uşor şi rapid de personal necalificat şi distribuirea acestora spre rezolvare către un grup necunoscut de utilizatori de pe Internet" - posibilitatea de a accesa un număr practic nelimitat de utilizatori face din aceasta un candidat ideal pentru evaluarea subiectivă.

În prezent domeniul de "crowd-sourcing" se stabileşte ca domeniu de sine stătător asociat metodelor de analiză multimedia. Tot mai multe studii dovedesc faptul că rezultatele obţinute în urma "crowd-sourcing" pot fi comparabile cu cele obţinute de utilizatori experţi [Nowak 10]. Totuşi sistemul de "crowd-sourcing" nu este perfect şi nu orice evaluare poate fi proiectată prin intermediul "crowd-sourcing".

Principala problemă este dată de controlul calităţii rezultatelor. Dacă în cazul campaniilor de evaluare utilizatorii sunt aleşi astfel încât să fie familiarizaţi cu domeniul precum şi să fie motivaţi în a furniza o evaluare de calitate (voluntar, în interes de cercetare, eventual remunerat), în cazul "crowd-sourcing" nu există un control direct asupra alegerii utilizatorilor iar calitatea rezultatelor nu poate fi controlată în mod direct, participanţii la studiu fiind motivaţi în principal de un câştig financiar asociat fiecărei sarcini de

rezolvat care este extrem de redus (exemplu 4$ pe oră). Din perspectiva organizării evaluării, singurul mecanism de creştere a calităţii evaluări este dat de modul de concepere al evaluării care trebuie să fie unul intuitiv, simplu, rapid şi atractiv pentru utilizator.

Dintre platformele de "crowd-sourcing" existente una dintre cele mai populare este Amazon Mechanical Turk[1]. Aceasta este totuşi limitată în a fi accesibilă doar pentru persoane ("requesters" - persoanele care formulează sarcinile ce trebuiesc rezolvate de utilizatori) care au coordonate bancare în Statele Unite. O alternativă la aceasta este platformă este Crowdflower[2]. Cererile de lucru create în Crowdflower pot fi publicate pe diverse canale de "crowd-sourcing" ce includ şi platforma Amazon Mechanical Turk.

În ceea ce priveşte controlul calităţii, există o serie de facilităţi care ţin mai mult de modul de alegere al utilizatorilor decât de evaluarea acestora. De exemplu, în cazul platformei Amazon Mechanical Turk se poate opta pentru a alege utilizatori din anumite locaţii geografice, alege utilizatori în funcţie de performanţa acestora dovedită în alte sarcini efectuate anterior (cel mai probabil în domenii complet diferite) sau pe baza numărului de sarcini realizate anterior. Există şi posibilitatea de refuzare a rezultatelor considerate nesatisfăcătoare fără a implica costuri suplimentare. În cazul platformei Crowdflower aceasta introduce conceptul de "gold units" prin care încearcă să elimine utilizatorii cu performanţe slabe precum şi posibilitatea de generare de răspunsuri automate sau aleatorii. Practic, utilizatorilor li se cere să răspundă la cel puţin 4 întrebări al căror răspuns este deja cunoscut de sistem şi doar în cazul în care obţin o precizie de minim 70% răspunsurile acestora la sarcina curentă de rezolvat sunt luate în calcul. În cazul platformei Crowdflower nu există posibilitatea de a refuza răspunsurile considerate ca fiind nerelevante.

Indiferent de modul de implicare al utilizatorilor în procesul de evaluare, acest mod de abordare presupune un anumit grad de subiectivitate. Persoane diferite pot percepe diferit anumite informaţii (vezi şi exemplul din Figura 8.1). Astfel, se pune problema găsirii unei modalităţi de evaluare a gradului de subiectivitate dintre evaluările furnizate de utilizatori, informaţie ce este de regulă furnizată impreună cu rezultatele obţinute.

Una dintre abordările cele mai frecvent folosite constă în evaluarea gradului de concordanţă dintre evaluările realizate de utilizatori diferiţi pentru aceleaşi date, ceea ce se numeşte "inter-annotator agreement". Prezentăm în continuare modul de calcul al coeficientului Kappa [Carletta 96] ce reprezintă

[1] `https://www.mturk.com/mturk`
[2] `http://crowdflower.com`

o măsură statistică a concordanţei dintre răspunsurile furnizate de utilizatori diferiţi. Spre deosebire de alte mărimi similare, coeficientul Kappa ia în calcul şi concordanţa rezultatelor obţinută din întâmplare (aleator).

Să considerăm cazul a doi utilizatori care evaluează un număr de N entităţi ca aparţinând a C categorii (categoriile considerate sunt complementare). De exemplu poate fi vorba de etichetarea a N imagini ca fiind relevante sau nerelevante ($C = 2$ în acest caz). În acest caz coeficientul Kappa este dat de relaţia următoare:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{8.1}$$

unde $Pr(a)$ reprezintă probabilitatea observată relativă de concordanţă între utilizatori iar $Pr(e)$ reprezintă probabilitatea ipotetică de concordanţă datorată întâmplării. Dacă răspunsurile utilizatorilor sunt în concordanţă completă atunci valoarea lui κ este 1 iar similar, dacă există o disconcordanţă totală între răspunsuri κ este 0. În realitate o valoare a lui κ superioară a 0.6 este considerată ca fiind perfectă.

Pentru exemplificare să considerăm următoarele date (sursă Wikipedia): avem la dispoziţie 50 de propuneri de proiecte de cercetare ce sunt evaluate fiecare de câte doi evaluatori (notaţi A şi respectiv B). Aceştia atribuie propunerilor categoria "da" sau "nu" (semnificând acceptarea acestora pentru finanţare sau nu). Presupunând că datele obţinute sunt cele prezentate în Tabelul 8.1 (numerele corespund numărului de proiecte pentru care evaluatorii au furnizat răspunsul da sau nu) atunci probabilităţile $Pr(a)$ şi $Pr(e)$ sunt estimate în felul următor:

- $Pr(a)$: evaluatorii A şi B au acordat impreună calficativul "da" pentru 20 de proiecte şi respectiv "nu" pentru 15 proiecte astfel că probabilitatea de concordanţă a răspunsurilor este $Pr(a) = (20 + 15)/50 = 0.7$;

Tabelul 8.1: Exemplu de calcul al coeficientului Kappa (sursă Wikipedia).

		B	B
		"da"	"nu"
A	"da"	20	5
A	"nu"	10	15

- $Pr(e)$: în acest caz se observă următoarele: evaluatorul A a răspuns "da" pentru 25 de proiecte şi "nu" tot pentru 25 ceea ce înseamnă că evaluatorul A răspunde cu "da" pentru 50% din cazuri. Similar, evaluatorul B a răspuns "da" pentru 30 de proiecte şi "nu" pentru 20

ceea ce înseamnă că evaluatorul B răspunde cu "da" pentru 60% din cazuri.

Probabilitatea ca cei doi evaluatori să răspundă cu "da" în mod aleator este $0.5 \cdot 0.6 = 0.3$ iar probabilitatea ca ambii să răspundă cu "nu" este $0.5 \cdot 0.4 = 0.2$. Astfel, per total probabilitatea de concordanţă aleatoare este $0.3 + 0.2 = 0.5$.

Aplicând relaţia anterioară obţinem în acest caz un coeficient Kappa de 0.4 care indică o concordanţă relativ scăzută a rezultatelor.

8.2 Evaluarea obiectivă

O altă abordare a problemei evaluării performanţei sistemelor de indexare după conţinut o constituie metodele de evaluare aşa zisă obiectivă. Acestea se bazează pe evaluarea performanţelor cuantificând erorile de căutare cu diverse măsuri statistice matematice. Pentru a putea evalua o măsură de eroare este necesară cunoaşterea apartenenţei datelor la clasele căutate (datele să fie etichetate) sau cu alte cuvinte "ground truth".

Având în vedere faptul că este practic imposibil să dispunem de "ground truth" în cazul unei baze de date dinamice (de exemplu de pe Internet) sau chiar de dimensiune semnificativă, lucru ce ar face procesul de căutare inutil atâta timp cât datele sunt deja cunoscute, validarea obiectivă se realizează preliminar folosind seturi de date de test. Sistemul se calibrează astfel pentru performanţă optimală folosind aceste baze de test urmând a fi implementat practic ulterior în contextul real. Pentru ca rezultatele unui astfel de proces de evaluare să fie relevante la scară reală, seturile de date folosite trebuie să fie reprezentative şi cât mai diverse.

Ca ordin de măsură, în contextul actual, bazele de test pentru sistemele de căutare dupa conţinut a imaginilor tind să conţină până la milioane de imagini în timp ce în contextul video acestea sunt de ordinul sutelor de mii. Principala limitare este dată de efortul necesar etichetării acestora ce presupune analiza lor manuală de către experţi umani. De exemplu, dacă dorim validarea unui sistem de căutare a secvenţelor de gol într-o bază video de înregistrări de fotbal, fiecare dintre secvenţe trebuie parcursă manual şi etichetate momentele de timp în care apar secvenţele căutate. Pe baza acestor date, rezultatele obţinute de sistemul de căutare automată pot fi comparate cu rezultatele ideale obţinute manual.

În literatura de specialitate există o multitudine de abordări propuse pentru evaluarea obiectivă a performanţelor, pentru o descriere exhaustivă a

acestora cititorul se poate raporta la [Manning 08]. În cele ce urmează vom detalia unele dintre abordările cele mai frecvent întâlnite.

8.2.1 Precision-Recall

Dacă analizăm problema căutării datelor din perspectiva unui sistem de clasificare (vezi exemplu Secţiune 4.2) şi anume, rezultatele obţinute în urma căutării corespund de fapt unei clasificări binare a datelor existente, acestea fiind etichetate fie ca aparţinând clasei obiectului căutat ("query", clasa A), fie ca aparţinând celorlalte clase existente (clasa B), atunci erorile de căutare pot fi sintetizate în modul următor (vezi Tabel 8.2):

- *tp* sau "true positive": reprezintă obţinerea unui rezultat corect şi anume obiectul returnat de sistem a fost prezis ca aparţinând clasei A (clasa căutată) acesta corespunzând şi în realitate clasei A;
- *fp* sau "false positive": reprezintă obţinerea unui rezultat fals şi anume obiectul returnat de sistem a fost prezis ca aparţinând clasei A dar în realitate acesta corespunde unui obiect din clasa B ceea ce conduce la o predicţie falsă;

Tabelul 8.2: Erori statistice în cazul clasificării datelor.

		clasa reală	
		clasa A	clasa B
clasa prezisă	clasa A	tp (true positive)	fp (false positive)
	clasa B	fn (false negative)	tn (true negative)

- *fn* sau "false negative": reprezintă obţinerea tot a unui rezultat fals şi anume sistemul a prezis că obiectul returnat aparţine clasei B în realitate acesta fiind din clasa A fapt ce conduce la o non-detecţie, obiectul A (din clasa căutată) fiind pierdut;
- *tn* sau "true negative": reprezintă prezicerea rezultatului ca fiind un obiect din clasa B în măsura în care acesta este în realitate tot din clasa B această situaţie fiind o confirmare a absenţei obiectului căutat de tip A.

Cu alte cuvinte, în urma căutării se pot obţine două situaţii de eroare: obiectul căutat este estimat eronat ca fiind un obiect din altă clasă, eroare cuantizată de raportul fp; şi respectiv obiectul căutat nu este găsit, situaţie cuantizată de raportul fn.

Pe baza acestor erori sunt definite măsurile de performanţă numite *precision* şi *recall* astfel:

$$precision = \frac{tp}{tp + fp}, \quad recall = \frac{tp}{tp + fn} \tag{8.2}$$

Definite în acest fel, *precision* este o măsură a falselor detecţii iar *recall* o măsură a non-detecţiilor. Plaja de valori a acestora se găseşte în intervalul $[0; 1]$ unde 1 reprezintă cazul ideal în care nu există nici o falsă detecţie ($fp = 0$) şi respectiv toate documentele existente în bază au fost găsite ($fn = 0$). Se poate observa faptul că valoarea $tp + fn$ este o constantă şi reprezintă numărul total de obiecte de tip A existente în baza de date (numărul celor identificate corect + numărul celor care nu au fost returnate).

Dacă analizăm problema căutării datelor din perspectiva unui sistem de indexare clasic în care rezultatele căutării sunt reprezentate în ordinea descrescătoare a relevanţei acestora relativ la obiectul căutat (vezi exemplu Secţiune 6) atunci modul de calcul al *precision* şi *recall* este un pic diferit. Diferenţa provine din faptul că evaluarea performanţei se realizează de această dată pe un set de rezultate ordonate şi care nu reprezintă neapărat toate documentele disponibile din baza de date (se pot returna doar o parte din acestea în urma căutării - de exemplu în cazul bazelor de date de pe Internet rezultatele căutări sunt limitate la un număr ce poate fi gestionat de utilizator).

În acest context, *precision* este o măsură a procentului din documentele returnate ce sunt relevante pentru obiectul căutat ("query"):

$$precision = \frac{|\{documente\ relevante\} \cap \{documente\ returnate\}|}{|\{documente\ returnate\}|} \tag{8.3}$$

unde operatorul $|.|$ returnează numărul de elemente ale unei mulţimi.

Similar, *recall* este o măsură a procentului de documentele relevante pentru obiectul căutat ce au fost returnate în urma căutării şi anume:

$$recall = \frac{|\{documente\ relevante\} \cap \{documente\ returnate\}|}{|\{documente\ relevante\}|} \tag{8.4}$$

Dat fiind faptul că aceste măsuri sunt evaluate pentru o anumită căutare particulară, pentru a obţine o măsură globală de performanţă de regulă se calculează valorile medii ale acestora pentru un anumit număr de căutări. Dacă baza de date este cunoscută, atunci se poate realiza o evaluare exhaustivă în care fiecare document din bază este folosit pentru a specifica cererea de căutare iar performanţa sistemului este estimată ca valoare medie pentru toate căutările efectuate.

8.2.2 F-measure

Având în vedere cele două situaţii de eroare ce trebuiesc luate în calcul pentru evaluarea performanţelor indexării şi anume numărul de false detecţii şi respectiv numărul de non-detecţii, se pune problema care dintre acestea este mai importantă. Astfel, de exemplu un sistem de indexare care furnizează *precision* de 95% şi *recall* de 80% este preferabil unui sistem ce furnizează *recall* de 95% şi respectiv *precision* de 80%? Cu alte cuvinte, care dintre cele două situaţii sunt mai dezavantajoase, un sistem în care rata de documente relevante returnate este mai mare (număr de false detecţii redus) iar numărul total de documente relevante returnate din numărul total existent în bază este mai mic (numărul de non-detecţii mai mare), sau situaţia inversă?

În realitate, răspunsul depinde strict de domeniul de aplicaţie. Figura 8.2 prezintă estimativ importanţa celor două măsuri pentru o serie de domenii de aplicaţie [Worring 12]. Astfel, dacă considerăm ca domeniu de aplicaţie căutarea datelor pe Internet atunci cel mai important parametru este *precision* deoarece se doreşte ca rezultatele căutării să fie cât mai precise. În acelaşi timp nu este la fel de important faptul că în urma căutării nu obţinem toate rezultatele relevante existente, în cazul Internetului acesta fiind un număr practic nelimitat, ci este suficientă obţinerea a unei submulţimi a acestora. Este cunoscut faptul că în practică în urma căutării într-un sistem "on-line" ne limităm de regulă în a analiza doar primele câteva zeci de rezultate.

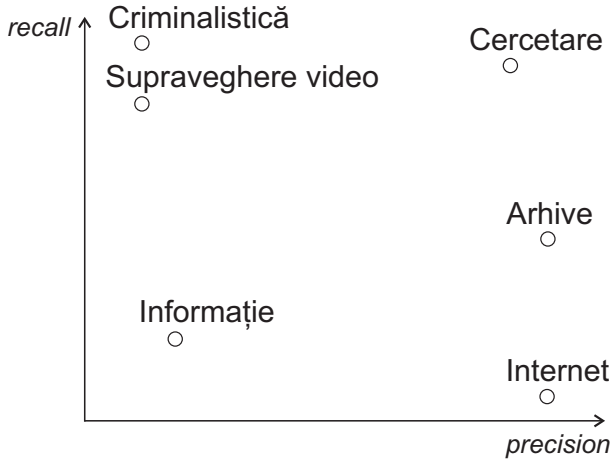


Figura 8.2: Gradul de importanţă al *precision* şi *recall* în funcţie de domeniul de aplicaţie (bazat pe informaţiile prezentate în [Worring 12]).

Pe de altă parte, dacă considerăm ca domeniu de aplicaţie un sistem spe-

cific expertizei criminalistice, de exemplu un sistem de identificare a amprentelor, în acest caz este mai important parametrul de *recall*. Cu alte cuvinte, este mai important ca sistemul să fie capabil să returneze toate documentele relevante existente în baza de date chiar dacă numărul de detecţii false este ridicat. Acestea pot fi reduse ulterior printr-o analiză manuală a rezultatelor dar absenţa unor documente relevante pentru căutare din rezultate nu mai poate fi corectată.

În acest context, în literatura de specialitate există un parametru care combină contribuţia celor două măsuri şi anume $F - measure$. Acesta este definit astfel:

$$F - measure = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall} \tag{8.5}$$

unde β reprezintă un parametru de reglaj al contribuţiei celor două măsuri. În funcţie de valoarea lui β, $F-measure$ poate evidenţia mai mult contribuţia uneia dintre cele două măsuri permiţând adaptarea evaluării la domeniul de aplicaţie.

Daca $\beta = 1$ atunci *precision* şi *recall* au ponderi egale ceea ce conduce la mărimea $F1 - score$ definită ca fiind media armonică dintre *precision* şi *recall*, astfel:

$$F1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{8.6}$$

8.2.3 Curbă de precision-recall şi ROC

Având în vedere faptul că de regulă sistemele de indexare returnează rezultatele în ordinea descrescătoare a relevanţei faţă de cererea de căutare ("ranking"), valorile estimate pentru *precision* şi *recall* sunt dependente de dimensiunea ferestrei de analiză a rezultatelor returnate. De exemplu, nu este acelaşi lucru dacă evaluăm performanţa pentru 100 de rezultate returnate sau pentru 200, în cazul din urmă fiind mai probabil ca numărul de rezultate corecte să fie mai mare.

Se pune astfel problema evaluării performanţei pentru puncte de operare ("operating points") diferite. Una dintre modalităţile cele mai frecvent folosite este accea de a reprezenta grafic *precision* în funcţie de *recall* pentru toată plaja de dimensiuni a ferestrei de rezultate până în punctul în care în aceasta se regăsesc toate datele căutate existente în baza de date.

Algoritmul de generare este următorul: pentru o anumită căutare în baza de date se consideră doar primele N_i rezultate obţinute, valoare astfel aleasă încât între acestea să se găsească exact i rezultate corecte. Valoarea lui i va varia de la 1 la $tp + fn$ (vezi ecuaţia 8.2), şi anume până în momentul în care

regăsim în fereastra considerată toate rezultatele corecte existente în baza de date.

În aceste condiţii, *precision* şi *recall* se evaluează în felul următor:

$$precision = \frac{1}{N_1},\ \frac{2}{N_2},\ ...,\ \frac{i}{N_i},\ ...,\ \frac{tp+fn}{N_{tp+fn}}$$
$$recall = \frac{1}{tp+fn},\ \frac{2}{tp+fn},\ ...,\ \frac{i}{tp+fn},\ ...,\ 1 \tag{8.7}$$

unde N_{tp+fn} reprezintă acea dimensiune a ferestrei pentru care obţinem toate rezultatele corecte existente în bază. Reprezentat în acest fel, graficul $precision - recall$ oferă o imagine asupra performanţei sistemului pentru toată plaja de puncte de operare, puntându-se stabili performanţa punctuală în oricare dintre acestea.

Figura 8.3 prezintă câteva exemple de grafice $precision - recall$ pentru un sistem perfect în care *precision* şi *recall* sunt 100%, un sistem complet ineficient în care *precision* şi *recall* sunt 0% şi un sistem real, sistemul propus în [Ionescu 13]. Primele două variante sunt variantele extreme, de performanţă maximă şi relativ minimă, în realitate performanţele sistemelor existente regăsindu-se între aceste două curbe (vezi Figura 8.3.(c)).

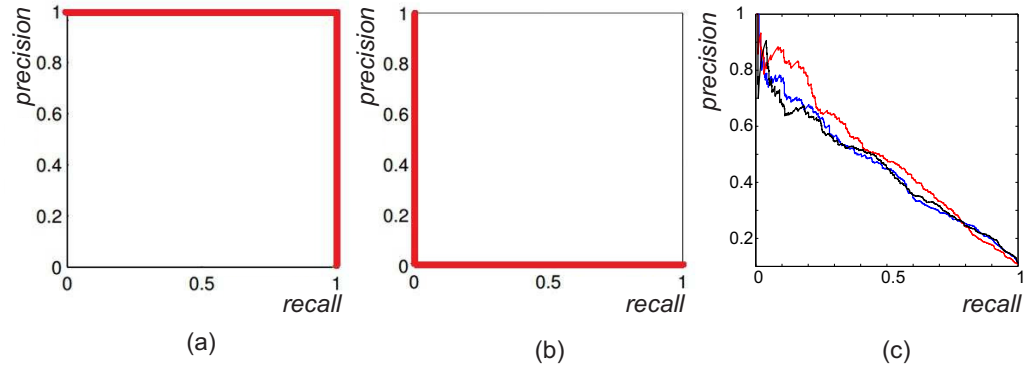


Figura 8.3: Exemple de grafice de tip $precision - recall$ pentru: (a) un sistem perfect, (b) un sistem complet ineficient, (c) un sistem real de căutare automată a segmentelor de violenţă din filme [Ionescu 13] (curbele sunt obţinute pentru diferite valori ale parametrilor sistemului).

O altă interpretare a graficului $precision - recall$ este aceea din perspectiva raportului de documente găsite corect (tpr) raportat la raportul de documente returnate eronat (fpr), ceea ce se numeşte curbă de tip Receiver Operational Characteristic sau ROC. Cele două rapoarte sunt definite în

modul următor:

$$tpr = \frac{tp}{tp + fn}, \quad fpr = \frac{fp}{fp + tn} \tag{8.8}$$

unde tp reprezintă numărul de documente returnate corect (vezi ecuaţia 8.2), fn reprezintă numărul de documente căutate care nu sunt returnate (non-detecţie), fp reprezintă numărul de documente fals detectate iar tn reprezintă numărul de documente ignorate (documente care sunt prezise corect ca neaparţinând clasei căutate). Definite în acest fel, tpr este o măsură a numărului de documente returnate corect iar fpr o măsură a numărului de documente returnate eronat.

Figura 8.4 prezintă două exemple de curbe ROC, în cazul unui sistem perfect în care tpr este 100% iar fpr este 0% cât şi în cazul unui sistem complet ineficient în care numărul de rezultate corecte este egal cu numărul de rezultate false, un astfel de sistem neputând fi practic utilizat. În realitate, pentru ca un sistem de indexare să ofere performanţe bune, curba ROC asociată trebuie să se situeze între cele două grafice, cât mai apropiată de sistemul ideal.

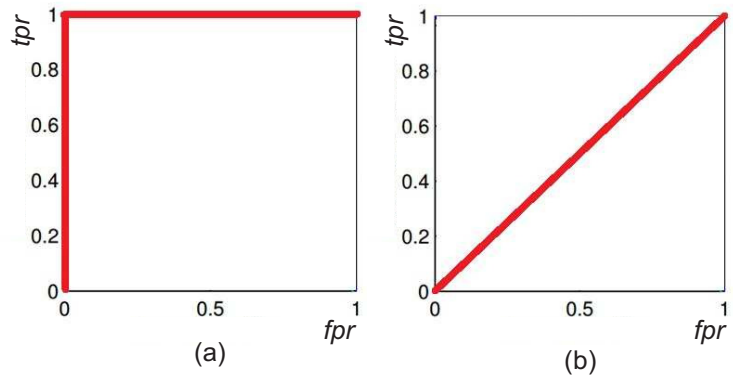


Figura 8.4: Exemple de grafice de tip ROC pentru: (a) un sistem perfect, (b) un sistem complet ineficient în care numărul de documente returnate eronat este egal cu numărul de documente returnate corect.

8.2.4 Mean Average Precision

În ultimii ani, pornind din contextul sistemelor de indexare video, s-a impus ca standard de evaluare a performanţelor sistemelor de indexare ceea ce numim Mean Average Precision sau MAP[3]. MAP furnizează o măsură a ca-

[3]vezi utilitar `http://trec.nist.gov/trec_eval`

lităţii sistemului pentru diferite valori ale *recall* (vezi ecuaţie 8.2), totul prin intermediul unei singure mărimi. Acesta se dovedeşte în practică a furniza o bună stabilitate şi discriminanţă în evaluarea diferitelor sisteme.

MAP este estimat în modul următor: dacă pentru o anumită cerere de căutare q_j, unde $j = 1, ..., |Q|$ cu Q reprezentând mulţimea căutărilor posibile pentru sistemul considerat (de exemplu, dacă sistemul permite indexarea în funcţie de obiecte atunci Q reprezintă mulţimea tuturor obiectelor din bază) iar operatorul $||.||$ returnează numărul de elemente ale unei mulţimi; mulţimea documentelor relevante din bază este $\{d_1, ..., d_{m_j}\}$ (numărul de documente relevante pentru q_j este m_j) iar R_{jk} reprezintă mulţimea primelor documente returnate până la documentul d_k (fereastra de rezultate care include şi documentul d_k), atunci MAP este definit ca:

$$MAP(Q) = \frac{1}{||Q||} \cdot \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} precision(R_{jk}) \tag{8.9}$$

unde *precision* este calculat aşa cum a fost definit în ecuaţia 8.2. Cu alte cuvinte, MAP reprezintă media *precision* pentru fereastra de rezultate ce include toate documentele relevante pentru o căutare (termenul *Average*), valoare ce este la rândul ei mediată pentru toate căutările posibile (termenul *Mean*). În cazul în care sistemul nu returnează nici un document relevant atunci MAP este 0%.

Pentru o singură cerere de căutare ("query") MAP poate fi aproximat ca fiind aria dintre graficului $precision-recall$ şi axa orizontală (vezi Figura 8.3) şi astfel pentru un set de căutări acesta va reprezenta aria medie a graficelor de $precision - recall$.

CAPITOLUL 9

Paradigme ale indexării

În capitolele anterioare am trecut în revistă punctual marea parte a problemelor de prelucrare aferente sistemelor de indexare automată după conţinut a datelor multimedia.

În acest ultim capitol vom face o trecere în revistă a barierelor tehnologice, principiale, ce trebuiesc depăşite pentru a putea soluţiona eficient problema căutării informaţiei. Acestea sunt enunţate în literatură sub denumirea de paradigme:

- **paradigma senzorială** ("sensor gap") reprezintă discrepanţa care există între informaţiile prezente în lumea reală 3D şi informaţiile înregistrate de senzori (de exemplu camere foto, video, microfoane, etc.), informaţii ce sunt folosite pentru analiza conţinutului datelor. De exemplu, în cazul imaginilor acestea nu sunt decât proiecţii plane 2D al lumii 3D. Mai mult, acelaşi obiect de interes poate conduce la un număr nelimitat de reprezentări diferite datorate perturbaţiei senzorilor sau a factorilor externi (vezi exemple din Figura 9.1). Astfel, o primă paradigmă ce trebuie depăşită este aceea a modelării informaţiei incomplete de care dispunem şi a variabilităţii acesteia. Practic metodele de analiză de conţinut încearcă să estimeze informaţiile lipsă, fie pe baza unor modele, sau prin compensarea cu informaţii suplimentare obţinute din alte surse;

- **paradigma semantică** ("semantic gap") reprezintă discrepanţa care există între informaţiile extrase în mod automat din date şi semnificaţia semantică pe care le-o putem atribuii acestora. Cu alte cuvinte, în ciuda

Figura 9.1: Un anumit obiect poate fi înregistrat sub o multitudine de reprezentări diferite datorate schimbării unghiului din care este reprezentat, schimbării de iluminare, schimbării fundalului sau ocluziei cu alte obiecte (sursă imagini [Snoek 10]).

faptului că un sistem poate funcţiona corect din punct de vedere al algoritmilor, şi chiar mai mult, poate fi antrenat să răspundă optimal pentru un anumit domeniu de aplicaţie sau set de date, în realitate rezultatele obţinute pot să nu corespundă aşteptărilor şi a modului de percepţie uman;

- **paradigma modelării** ("model gap") reprezintă imposibilitatea de a determina un model general pentru toate obiectele sau entităţile informaţionale existente în lume fiind limitaţi în a modela cazuri particulare, precum obiecte, concepte, evenimente şi aşa mai departe. Diversitatea informaţională existentă face imposibilă acoperirea tuturor cazurilor posibile;

Figura 9.2: Există o multitudine de obiecte şi concepte ce trebuiesc modelate pentru a putea fi accesate la nivel de informaţie.

- **paradigma intenţiei** ("intention/query gap") reprezintă discrepanţa dintre informaţiile pe care utilizatorul doreşte să le găsescă şi modul de exprimare a criteriilor de căutare într-un sistem de indexare (vezi Figura 9.3). Cele mai performante metode existente permit specificarea criteriilor de căutare sub fomă textuală. Acest mod de reprezentare este limitat la un număr redus de informaţii ce pot fi furnizate (de regulă cel mult o propoziţie) nereflectând în totalitate informaţia reală dorită;

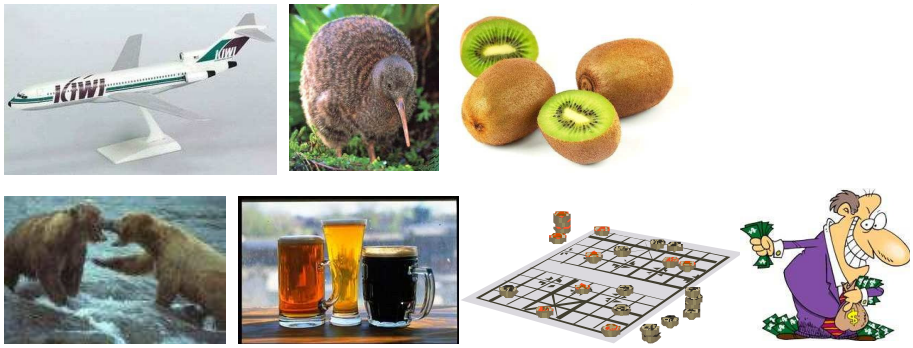


Figura 9.3: Există o multitudine de "întrebuinţări" ale aceluiaşi concept, de exemplu "kiwi" poate reprezenta atât o companie aeriană, un fruct sau o pasăre, "bear" (urs) este foarte similar cu "beer" (bere) sau "grid" (caroiaj) cu "greed" (lacom) (exemplu din cursul Indexarea Conţinutului Vizual, Constantin Vertan, Universitatea Politehnica din Bucureşti).

- **paradigma utilităţii** ("utility gap") reprezintă discrepanţa care există între rezultatele furnizate de sistem şi utilitatea reală practică a acestora pentru utilizator. Ca şi în cazul paradigmei semantice, sistemul poate fi performant şi să returneze utilizatorului o multitudine de informaţii relevante relativ la datele căutate, dar câte dintre aceste informaţii vor servi în mod real util utilizatorului.

Bibliografie

[Bimbo 99] A. Del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann Publishers, San Francisco, USA 1999.

[Bovik 09] Alan C. Bovik. *The Essential Guide to Video Processing*. Academic Press, ISBN: 978-0-12-374456-2, 2009.

[Carletta 96] J. Carletta. *Assessing agreement on classification tasks: The kappa statistic*. Computational Linguistics, vol. 22, nr. 2, pag. 249–254, 1996.

[CITIA 13] CITIA. *City of Moving Images, International Animated Film Festival of Annecy, France*. http://www.citia.info, 2013.

[Ciuc 05] M. Ciuc & C. Vertan. *Prelucrarea Statistică a Semnalelor*. Editura MatrixRom, `http://www.miv.ro/books/MCiuc_CVertan_PSS.pdf`, 2005.

[Deza 06] E. Deza & M.M. Deza. *Dictionary of Distances*. Elsevier Science, 1st edition, ISBN-10:0444520872, 2006.

[Flickner 95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Patkovic, D. Steele & P. Yanker. *Query by Image and Video Content: The QBIC System*. IEEE Computer, vol. 28, nr. 9, pag. 23–32, septembrie 1995.

[Gauglitz 11] S. Gauglitz, T. Höllerer & M. Turk. *Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking*. Int J. Comput Vis, vol. DOI 10.1007/s11263-011-0431-5, 2011.

[Gómez-Pérez 04] A. Gómez-Pérez, M. Fernández-López & O. Corcho. *Lecture Notes: Multimedia Information Systems*. Ontological Engineering: With Examples from the Areas of Knowledge Management, E-commerce and the Semantic Web, Springer. ISBN 978-1-85233-551-9., 2004.

[Ionescu 09] B. Ionescu. *Analiza şi Prelucrarea Secvenţelor Video: Indexarea Automată după Conţinut*. Editura Tehnică Bucureşti, ISBN 978-973-31-2354-5, 2009.

[Ionescu 10] B. Ionescu, L. Ott, P. Lambert, D. Coquin, A. Pacureanu & V. Buzuloiu. *Tackling Action - Based Video Abstraction of Animated Movies for Video Browsing*. SPIE - Journal of Electronic Imaging, vol. 19, nr. 3, 2010.

[Ionescu 11] B. Ionescu, C. Rasche, C. Vertan & P. Lambert. *A Contour-Color-Action Approach to Automatic Classification of Several Common Video Genres*. Springer-Verlag LNCS - Lecture Notes in Computer Science, Eds. M. Detyniecki, P. Knees, A. Nurnberger, M. Schedl and S. Stober, vol. 6817, pag. 74–88, 2011.

[Ionescu 12a] B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan & P. Lambert. *Content-based Video Description for Automatic Video Genre Categorization*. International Conference on MultiMedia Modeling, 2012.

[Ionescu 12b] B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan & P. Lambert. *Video Genre Categorization and Representation using Audio-Visual Information*. SPIE - Journal of Electronic Imaging, vol. 21, nr. 2, 2012.

[Ionescu 13] B. Ionescu, J. Schlüter, I. Mironică & M. Schedl. *A Naive Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies*. ACM International Conference on Multimedia Retrieval, 2013.

[Jain 89] Anil K. Jain. Fundamentals of digital image processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[Kelly 03] D. Kelly & J. Teevan. *Implicit Feedback for Inferring User Preference: a Bibliography*. International Conference on Research and Development in Information Retrieval, vol. 37, nr. 2, pag. 18–28, 2003.

[Knees 07] P. Knees, M. Schedl, T. Pohle & G. Widmer. *Exploring Music Collections in Virtual Landscapes*. IEEE MultiMedia, vol. 14, nr. 3, pag. 46–54, 2007.

[Knees 09] P. Knees, T. Pohle, M. Schedl, D. Schnitzer, K. Seyerlehner & G. Widmer. *Augmenting Text-Based Music Retrieval with Audio Similarity*. International Society for Music Information Retrieval, 2009.

[Kotsiantis 07] S.B. Kotsiantis. *Supervised Machine Learning: A Review of Classification Techniques*. Informatica, vol. 31, pag. 249–268, 2007.

[Kyungpook 06] National University Kyungpook. *Artificial Intelligence Laboratory*. `http://ailab.kyungpook.ac.kr/vindex/video-view.html`, 2006.

[Lamel 08] L. Lamel & J.-L. Gauvain. *Speech Processing for Audio Indexing*. Int. Conf. on Natural Language Processing, vol. LNCS, 5221, pag. 4–15, 2008.

[Lan 12] Z. Lan, L. Bao, S.-I. Yu, W. Liu & A.G. Hauptmann. *Double Fusion for Multimedia Event Detection*. International Conference on Multimedia Modeling, Klagenfurt, Austria, 2012.

[Larson 10] Ray R. Larson. *Blind Relevance Feedback for the ImageCLEF Wikipedia Retrieval Task*. CLEF 2010 LABs and Workshops, Notebook Papers, pag. 22–23, 2010.

[Lienhart 01] R. Lienhart. *Reliable Transition Detection in Videos: A Survey and Practitiner's Guide*. MRL, Intel Corporation, `http://www.lienhart.de/Publications/IJIG_AUG2001.pdf`, august, Santa Clara, USA 2001.

[Maillet 03] S.M. Maillet. *Content-Based Video Retrieval: An Overview*. `http://viper.unige.ch/~marchand/CBVR/`, 2003.

[Manning 08] C.D. Manning, P. Raghavan & H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, `http://nlp.stanford.edu/IR-book/`, 2008.

[Mathieu 10] B. Mathieu, S. Essid, T. Fillon, J. Prado & G. Richard. *YAAFE an Easy to Use and Efficient Audio Feature Extraction Software*. 11th ISMIR conference, Utrecht, Netherlands, 2010.

[Mingqiang 08] Y. Mingqiang, K. Kidiyo & R. Joseph. *A Survey of Shape Feature Extraction Techniques*. Pattern Recognition, pag. 43–90, 2008.

[Mironică 12a] I. Mironică, B. Ionescu & C. Vertan. *Hierarchical Clustering Relevance Feedback for Content-Based Image Retrieval*. 10th International Workshop on Content-Based Multimedia Indexing, Annecy, France 2012.

[Mironică 12b] I. Mironică, B. Ionescu & C. Vertan. *The Influence of the Similarity Measure to Relevance Feedback*. 20th European Signal Processing Conference EUSIPCO, 2012.

[Nakazato 01] M. Nakazato & S. T. Huang. *3D MARS: Immersive virtual reality for content based image retrieval*. IEEE International Conference on Multimedia and Exposition, pag. 45–48, 2001.

[Nowak 10] S. Nowak & S. Rüger. *How reliable are annotations via crowdsourcing? a study about inter-annotator agreement for multi-label image annotation*. Int. Conf. on Multimedia Information Retrieval, pag. 557, 2010.

[Orchard 91] M. Orchard & C. Bouman. *Color Quantization of Images*. IEEE Trans. on Sig. Proc., vol. 39, nr. 12, pag. 2677–2690, 1991.

[Over 12] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Barbara Shaw, Wessel Kraaij, Alan F. Smeaton & Georges Quéenot. *TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics*. In Proceedings of TRECVID 2012. NIST, USA, 2012.

[Reynertson 70] A. J. Reynertson. *The Work of the Film Director*. Hastings House, 1970.

[Rocchio 71] J. Rocchio. *Relevance Feedback in Information Retrieval*. The Smart Retrieval System – Experiments in Automatic Document Processing, Prentice Hall, Englewood Cliffs NJ, pag. 313–323, 1971.

[Rooij 08] O. Rooij, C. G. M. Snoek, & M. Worring. *Mediamill: Fast and effective video search using the ForkBrowser*. ACM International Conference on Image and Video Retrieval, 2008.

[Rooij 10] O. Rooij, M. Worring & J. J. van Wijk. *MediaTable: Interactive Categorization of Multimedia Collections*. IEEE Computer Graphics and Applications, vol. 30, nr. 5, pag. 42–51, 2010.

[Rubner 00] Y. Rubner, C. Tomasi & L.J. Guibas. *The Earth Mover's Distance as a Metric for Image Retrieval*. International Journal of Computer Vision, vol. 40, nr. 2, pag. 99–121, 2000.

[Rui 99] Y. Rui, T. Huang & S.-F. Chang. *Image Retrieval: Current Techniques, Promising Directions and Open Issues*. Journal of Visual Communication and Image Representation, vol. 10, nr. 1, pag. 39–62, 1999.

[Schoeffmann 11] K. Schoeffmann & L. Boeszoermenyi. *Image and Video Browsing with a Cylindrical 3D Storyboard*. ACM International Conference on Multimedia Retrieval, 2011.

[Seyerlehner 10] K. Seyerlehner, M. Schedl, T. Pohle & P. Knees. *Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation*. 6th Annual Music Information Retrieval Evaluation eXchange (MIREX-10), Utrecht, Netherlands, 2010.

[Shirahama 11] K. Shirahama & K. Uehara. *Query by Virtual Example: Video Retrieval Using Example Shots Created by Virtual Reality Techniques*. Sixth International Conference on Image and Graphics, pag. 829–834, 2011.

[Smeulders 00] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta & R. Jain. *Content-Based Image Retrieval at the End of the Early Years*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, nr. 12, pag. 1349–1380, decembrie 2000.

[Snoek 05] C. G. M. Snoek, M. Worring & A. W. M. Smeulders. *Early versus Late Fusion in Semantic Video Analysis*. ACM Multimedia, 2005.

[Snoek 10] C.G.M. Snoek & A.W.M. Smeulders. *Video Search Engines*. IEEE Conference on Computer Vision and Pattern Recognition, `http://staff.science.uva.nl/~cgmsnoek/videosearch2010/`, 2010.

[Stöttinger 10] Julian Stöttinger, Bogdan Tudor Goras, Nicu Sebe & Allan Hanbury. *Behavior and properties of spatio-temporal local features under visual transformations*. 2010.

[Trémeau 04] A. Trémeau, C. Fernandez-Maloigne & P. Bonton. *Image Numérique Couleur: De l'Acquisition au Traitement*. DUNOD ISBN 2-10-006843-1, 2004.

[Truong 07] B.T. Truong & S. Venkatesh. *Video Abstraction: A Systematic Review and Classification*. ACM Transactions on Multimedia Computing, Communications and Applications, vol. 3, nr. 1, 2007.

[Tuceryan 93] M. Tuceryan & A. K. Jain. *Texture analysis*. The Handbook of Pattern Recognition and Computer Vision (2nd Edition), pag. 235–276, 1993.

[Wallach 06] Hanna M. Wallach. *Topic Modeling: Beyond Bag-of-Words*. University of Cambridge, `https://people.cs.umass.edu/~wallach/talks/beyond_bag-of-words.pdf`, 2006.

[Welling 05] M. Welling. *Support Vector Machines*. Note de curs, University of Toronto, Department of Computer Science, Canada, `http://www.ics.uci.edu/~welling/classnotes/papers_class/SVM.pdf`, 2005.

[Witten 05] I.H. Witten & E. Frank. *Data Minning - Practical Machine Learning Tools and Techniques*. Elsevier, Morgan Kaufman Publishers, second edition, pag. 265–270, 2005.

[Worring 03] M. Worring. *Lecture Notes: Multimedia Information Systems*. Intelligent Sensory Information Systems, University of Amsterdam, 2003.

[Worring 12] M. Worring. *Multimedia Analytics: Exploration of Large Multimedia Collections*. keynote la International Workshop on Content-Based Multimedia Indexing, `http://www.polytech.univ-savoie.fr/fileadmin/polytech_autres_sites/sites/cbmi2012/templates/fichiers/cbmi2012-worring.pdf`, 2012.