



Opleiding Informatie- en
Bibliotheekwetenschap

Kritische Bespreking voor het
vak Structurering van
Informatie

Prof. E. De Smet
Cursist Patricia Stegen



Inhoudsopgave

Inleiding	2
Wie of wat is YouTube?	3
Het concept 'You' en Web 2.0	3
Structuur en navigatie van YouTube	4
Structuur	4
Hoe navigeren doorheen You Tube?	5
Indexering en <i>Information Retrieval</i>	6
Zoeken in YouTube	6
Een experiment	7
De theorie	8
Technieken.....	8
Conclusie	11
Lijst met afbeeldingen	11

Inleiding

Structurering van Informatie heeft als finaliteit het vindbaar maken van relevante informatie.

Informatieve teksten worden opgeslagen in databanken, waardoor ze eenvoudig toegankelijk zijn. Maar wat met bewegend beeldmateriaal? Hoe moet dit ontsloten worden om relevante zoekresultaten te verkrijgen?

Ik besloot dit te gaan bekijken vanuit de invalshoek van één van de meest bekende spelers van participatieve media in de hedendaagse webomgeving: 'YouTube'.

Wie of wat is YouTube?

YouTube is opgericht in februari 2005 en is 's werelds meest populaire online videocommunity, die het miljoenen mensen mogelijk maakt oorspronkelijke video's te ontdekken, te bekijken en te delen. YouTube biedt een forum dat iedereen kan gebruiken om wereldwijd contact te leggen met anderen, mensen te informeren en te inspireren. De site fungeert als een distributieplatform voor makers van oorspronkelijke inhoud en voor adverteerders, hoe groot of klein dan ook. In november 2006, een jaar na de start ervan, werd YouTube overgenomen door Google Inc. en ging als één van de meest spraakmakende acquisities ooit de geschiedenis in.

(‘YouTube Fact Sheet’)¹

Kortom, YouTube is het eerste onvervalste en populaire massa-platform waar zowel amateurs als professionelen hun zelfgemaakte video's kunnen delen met de rest van de wereld. Een platform dat groot is geworden door het harde werk van haar gebruikers, want het zijn de zogenaamde *users* van YouTube die video's uploaden, teksten schrijven en berichten achterlaten bij videobeelden van andere gebruikers. YouTube wordt dan ook een toonaangevend voorbeeld genoemd van een platform dat bestaat uit door de gebruikers aangemaakte inhoud (*user generated content*).

Met dit gegeven ben ik op zoek gegaan naar informatie op het internet en in de databank 'Web of Science', lees verder WoS.

Het concept 'You' en Web 2.0

Ik begon mijn zoektocht naar een definitie van het begrip *user generated content*. Dit leverde bij 'Google Scholar' 906 000 resultaten op. Ironisch genoeg is de bovenste link een verwijzing naar 'Wikipedia'. Ik zeg ironisch, omdat een definitie niet te vinden is op de klassieke door een redactie ontwikkelde 'Encyclopedia Britannica Online'². Maar ook en vooral omdat 'Wikipedia' een voorbeeld is van hoe *user generated content* wordt gehanteerd op het internet, heel populair is, maar door vele wetenschappers ook verworpen wordt.

Ik zet mijn zoektocht naar een definitie verder op 'Freebase'³, dat heel wat minder resultaten (100) oplevert. Het begrip wordt hier door een aantal personen omschreven als: *de productie van inhoud door het grote publiek in plaats van door betaalde professionals en deskundigen in het veld. Het wordt gebruikt voor een breed scala van toepassingen, zoals vraag-antwoord databanken, digitale video, blogging, podcasting, mobiele telefoon fotografie en is meestal te vinden op het internet via blogs en wiki's. Consument-generated media en user created content worden aangehaald als synoniemen.*

Ook hier wordt weer verwezen naar 'Wikipedia'.

Bij het zoeken in WoS heb ik het anders aangepakt. Een zoekactie met de term 'YouTube' leverde 223 resultaten op, hetgeen ik verfijnd heb met '*user generated content*' dat me 28 resultaten opleverde, waarvan ik er 5 selecteerde om verder opzoekingen te doen in 'Google Scholar'.

¹ http://www.youtube.com/t/fact_sheet - geraadpleegd op 20091212

² <http://www.britannica.com/> - geraadpleegd op 20091212

³ <http://www.freebase.com/> - geraadpleegd op 20091220

Zodoende las ik in een rapport van de 'Organisation for Economic Co-operation and Development'⁴ dat aangezien er geen algemeen aanvaarde definitie van *user generated content* voorhanden is, zij de volgende definitie hanteren:

Inhoud openbaar gemaakt via het internet, die een 'zekere mate van creatieve inspanning' weerspiegelt, en die ontstaan is 'buiten de professionele routines en praktijken'. In het rapport staat dat het gebruik van het internet wordt gekenmerkt door meer participatie en interactie van Internetgebruikers, die het www gebruiken om te communiceren en zich te uiten. Het concept krijgt de naam: 'het participatieve web', dat ontstaan is als de ontwikkeling die inherent is aan de mogelijkheden om het internet te gebruiken op grotere schaal, waarmee ze verwijzen naar Web 2.0.

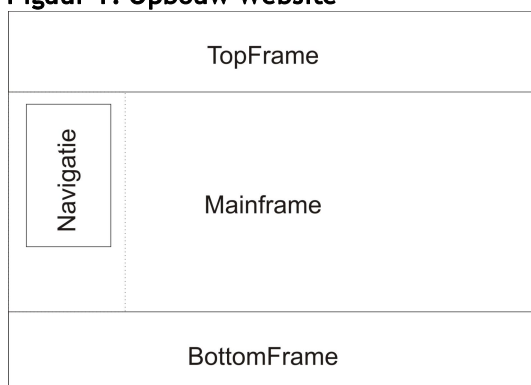
Structuur en navigatie van YouTube

Voor het aanmaken van een account hou je er best rekening mee dat gebruikersnamen niet hoofdlettergevoelig zijn.

Let op dat je de eerste keer meteen de juiste keuze maakt, kies bijvoorbeeld een naam die niet te lang is, want het is niet mogelijk om je gebruikersnaam nadien nog te veranderen. Je kan enkel een nieuwe 'account' aanmaken, maar je merknaam, die kan dienen als handelsmerk, ben je dan wel kwijt. Met één en hetzelfde e-mailadres kan je overigens zoveel accounts aanmaken als je maar wil.⁵

Structuur

Figuur 1: Opbouw Website



De website is opgebouwd uit een frameset met een TopFrame, een MainFrame en een BottomFrame.

Het MainFrame is nog eens opgedeeld in twee secties: een navigatiekolom en een gedeelte voor de inhoud.

Figuur 2: in het TopFrame staat het logo met een zoekvenster en een aantal hyperlinks.



De hyperlinks aan de rechterkant zijn voor gebruikers die een account willen aanmaken of inloggen op een bestaande account. Op de pagina 'Geschiedenis' kan je de volledige historie van je account bekijken.

⁴ Wunsch-Vincent, S.; Vickery, G. - *Participative Web: User-created content* - 2007 - bron: <http://www.oecd.org/dataoecd/57/14/38393115.pdf> - geraadpleegd op 20091220

⁵ Burgess, Jean E. and Green, Joshua B. - *YouTube : online video and participatory culture* - Digital Media & Society Polity Press - Cambridge, 2009

De belangrijkste hyperlink is 'Account maken'. Zonder een account of beter gezegd 'channel', lees verder kanaal, is het wel mogelijk video's te bekijken, maar kan je niets toevoegen.

Opmerkelijk is dat de pagina's 'Aanmelden', 'Abonnementen' en 'Uploaden' precies dezelfde inhoud hebben.

De website heeft een navigatiekolom aan de linkerkant van de pagina. Van hieruit word je doorverwezen naar de onderliggende pagina's, die verdeeld zijn in de volgende rubrieken: Categorieën, Shows, Movies, Wedstrijden en Evenementen. De optie Video's heeft nog een extra rubriek: Aanbevolen voor U. De eerste 3 rubrieken zijn op hun beurt nog eens opgedeeld in een hele reeks toepasselijke thema's. Elk thema is een link naar nog een aparte webpagina.

Hoe navigeren doorheen You Tube?

Er zijn twee belangrijke zoekingen: Video's en Kanalen; beide ingangen zijn onderverdeeld in categorieën.

Categorieën

Elke pagina bevat een lijst met *thumbnails*⁶ van video's voorzien van een beschrijving die titel (en ondertitel), aantal keren bekeken en uitvoerder vermeld.

Als je op een *thumbnail* parkeert, krijg je (meestal) aan de hand van een aantal beelden, lees verder *shots*, een samenvatting van de betreffende video te zien. Dit geeft al een goede indruk van de inhoud van het filmpje.

Vanaf de pagina 'Meest bekeken' is er de mogelijkheid om te verfijnen op 'Populairst' en 'HD'.

Shows

Vanaf deze pagina's kan je verder navigeren naar de volgende onderliggende pagina's: Alfabetisch (bevat een index), Nieuwste, Populair en Netwerk.

De beschrijving bestaat uit: titel, thema, aantal afleveringen (en clips) en een korte beschrijving van de inhoud.

'On the fly' opent zich een venster met extra metadata: titel, uitvoerder(s) en een uitgebreidere beschrijving van de inhoud.

Movies

Is op dezelfde manier opgezet dan Shows, met het verschil dat er in plaats van Netwerk, van Studio wordt gesproken.

De beschrijving bestaat uit: titel, het thema en het aantal keren bekeken.

Wedstrijden

Deze pagina is onderverdeeld in Wedstrijden (die nog lopende zijn) en Vorige wedstrijden.

De beschrijving bestaat uit: enkel de titel.

Evenementen

Deze pagina bevat een apart zoekvenster waar je kan zoeken op: 'Vindt evenementen'.

De beschrijving bestaat uit: titel, datum, plaats en een korte omschrijving.

⁶ Een verkleinde weergave van een foto of afbeelding, vaak gebruikt in websites

Indexering en Information Retrieval

Vanuit de bibliotheekwereld zijn we vertrouwd met het indexeren van bewegend beeld in de vorm van het globaal ontsluiten van dvd's die tot de collectie behoren. Deze materialen verschillen immers wezenlijk niet veel van boeken. De materialen hebben eveneens een cover waar allerlei gegevens uit af te leiden zijn, zoals titel, namen van medewerkers, datum van uitgave, enz.

Maar wat wanneer men te maken heeft met onafgewerkt materiaal? Het film- en videoarchief van de VRT bijvoorbeeld, beschrijft dit soort materiaal zoals andere archieven hun materiaal beschrijven: ze geven **manueel** een globale beschrijving van het voorwerp.

Gebruikers van YouTube doen in het feite precies hetzelfde: ze kennen een aantal trefwoorden toe waarmee ze hun filmpje willen typeren. Commerciële organisaties doen dat dan weer veel gedetailleerder, zodat de klant zeer gericht kan zoeken naar materiaal. Deze manier van werken heeft het grote voordeel dat het goed gedaan is en er bijgevolg weinig ruis zal zijn bij het zoeken. Het grote nadeel is dat het heel tijdrovend, subjectief en niet specifiek genoeg is.

Grote televisiestations beschrijven hun materiaal dan weer op het niveau van afzonderlijke beelden die ononderbroken opgenomen werden, zodat ze die later kunnen terugvinden voor hergebruik.

De toegang tot de gewenste stukjes informatie op een efficiënte manier ontsluiten wordt echter enorm bemoeilijkt door de enorme hoeveelheid en de diversiteit van het gepubliceerde video materiaal.

Daarom organiseert YouTube video's en de *retrieval* ervan via metadata, zoals de titel van de video, trefwoorden en beschrijvingen die door de gebruiker toegekend zijn. *Zo ontstaat een heel archief van beschrijvingen. De indeling die de gebruikers maken vormen een soort van taxonomie (classificatie). Omdat die door het gewone volk gemaakt wordt, spreekt men van folksonomy.*⁷

Zoeken in YouTube

Een zoekactie gebeurt op de volledige inhoud (cfr. *full text*) van de video en op de metadata. De standaarden MPEG-7 en MPEG-21 gaan over metadata.

De Moving Picture Experts Group (MPEG) heeft meerdere standaarden met betrekking tot het beschrijven van audio- en videomateriaal.⁸ Ik wou dit even vermelden, maar ga hier niet verder op in. Het zou een onderwerp voor een paper op zich kunnen zijn.

De gebruiker krijgt de mogelijkheid om geavanceerd te zoeken:

- Al deze woorden;
- Deze exacte zin;
- Één of meer van deze woorden;
- Geen van deze woorden.

Er is ook de mogelijkheid om je zoekactie te verfijnen met de volgende opties: Resultaattype, Sorteren op, Datum *Upload*, Categorieën, Duur, Kenmerken en Bevat.

⁷ De Keyser P. - *Cursus inhoudelijke documentbeschrijving: indexsystemen* - Genk, 2007

⁸ <http://www.chiariglione.org/mpeg/>, geraadpleegd op 20091212

Een experiment

Zoekwoorden kunnen in verschillende talen worden ingevoerd.

Figuur 3: zoekactie met als zoekwoord 'fótbolti', d.w.z. 'voetbal' in het IJslands

The screenshot shows a YouTube search page for the term 'fótbolti'. The search bar contains 'fótbolti' and the search button is visible. The page displays search results for 'fótbolti', showing 1-20 of about 243 results. The main content area is divided into two columns. The left column features a list of videos, including 'Fótbolti 6 fl KA KR' (Goðamót 2009) and 'Övissuferð Fótbolti'. Below this is a 'Playlist Results for fótbolti' section with two playlists: 'Fótbolti 17 videos' and 'Fótbolti 13 videos'. The right column is titled 'Featured Videos' and includes videos like 'fyndin fótbolti og...', 'Uppskeruhátíð BÍ', 'Video fyrir leik KV og Ýmis', 'Bloopersypa', and 'FH Mafian í Køben'. Each video entry includes a thumbnail, title, duration, and view count.

Deze zoekactie geeft 243 treffers.

Figuur 4: dezelfde zoekactie met als zoekwoord 'bóng đá' (Vietnamese)

The screenshot shows a YouTube search page for the term 'bóng đá'. The search bar contains 'bóng đá' and the search button is visible. The page displays search results for 'bóng đá', showing 1-20 of about 4,760 results. The main content area is divided into two columns. The left column features a list of videos, including 'Trận bóng đá đặc biệt' and 'clip hài đá bóng'. Below this is a 'Playlist Results for bóng đá' section with two playlists: 'Bóng đá - Teaching 66 videos' and 'amazing football freestyle 2009-bóng đá duong pho nghe thuật 157 videos'. The right column is titled 'Featured Videos' and includes videos like 'Chung kết bóng đá...', 'Khi cầu thủ bỏ bóng đá...', 'Bóng đá hà...', 'siêu bóng đá Clip vn', and 'Cuồng nhiệt bóng đá'. Each video entry includes a thumbnail, title, duration, and view count.

Deze zoekactie leverde 4 760 items en een heel ander resultaat op.

Ik deed nog een aantal zoekacties in andere vreemde talen, zoals ‘soka’ (Swahili): 5100, ‘sopakbola’ (Indonesië): 609, ‘pel-droed’ (Wales): 104 resultaten en een zoekactie in het Engels die 693 000 items opleverde.

Hieruit kunnen we concluderen dat het zoekresultaat sterk beïnvloed wordt door de taal.

Na mijn experimentje stelde ik me dan ook de vraag: ‘hoe gaat het brein van YouTube te werk om tegemoet te komen aan de vragen van de gebruiker?’

In een aantal artikels die ik uit Wos had gehaald, werd het één en ander verduidelijkt.

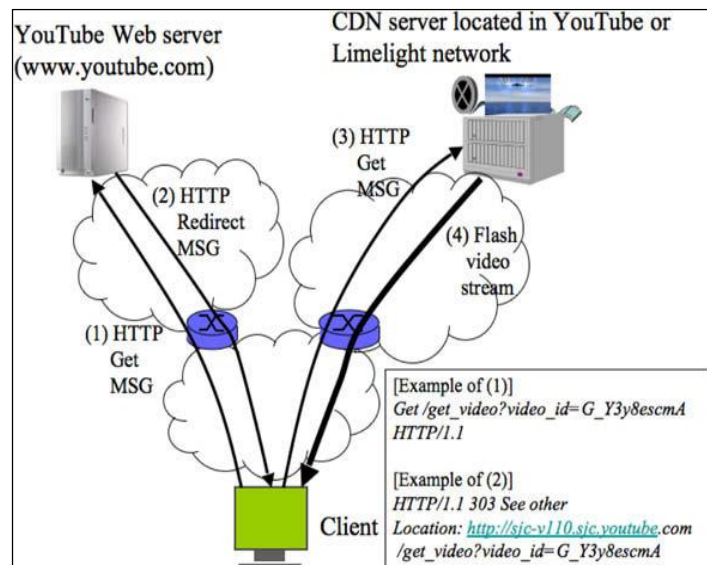
De theorie

We zien een illustratie van de communicatie tussen de klant, de YouTube-server en een server van het *Content Distribution Network (CDN)*. Wanneer een klant een bepaalde video heeft gekozen, wordt een ‘HTTP GET-bericht’ verstuurd vanaf de client naar de YouTube webserver.

Dit bericht geeft aan dat de klant een bepaalde video vraagt die wordt geïdentificeerd door een unieke identificatiecode-video, in dit voorbeeld *G_Y3y8escmA*.

Na ontvangst van het GET-bericht, antwoordt de web-server met een HTTP 303-Zie ander bericht’, namelijk het bericht via omleiding. Dit bericht bevat een locatie response-header veld, dat de klant doorverwijst naar de video-server waarop de video wordt gestreamd.⁹

Figuur 5: Video retrieval in YouTube



Aan de hand van resultaatmetingen, die in het artikel uitvoerig worden besproken, is het mogelijk om met ingewikkelde wiskundige formules te berekenen welke video’s met welke frequentie worden opgevraagd. Dankzij dergelijke onderzoeken is het mogelijk om technieken te bepalen om *information retrieval* van videomaterialen te verbeteren en zo de verhouding opbrengst/precisie te kunnen verhogen.

Technieken

Het automatisch toekennen van metadata gebeurt aan de hand van een grote hoeveelheid video’s over één en hetzelfde onderwerp. Het trefwoord wordt toegekend op basis van inhoudelijke kenmerken van het beeldmateriaal. Dit wordt inhoudelijk georiënteerde indexering (*content-based-indexing*) genoemd.

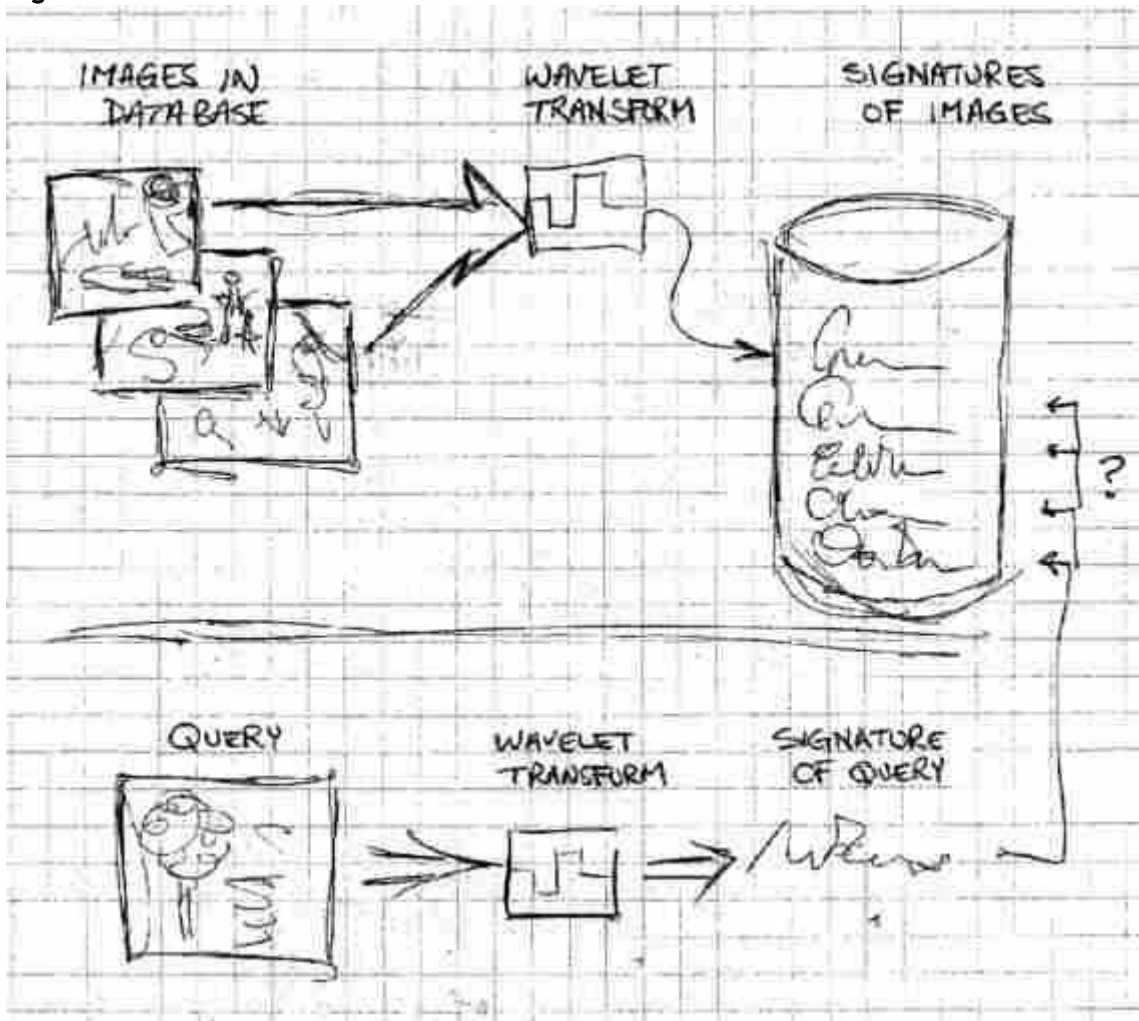
The Wavelet Transformation

Deze toepassing van Christian Langreiter is gebaseerd op het werk van drie onderzoekers van de University of Washington in 1995: Chuck Jacobs, Adam

⁹ Zink M, Suh K, Gu Y, et al. - *Characteristics of YouTube network traffic at a campus network / Measurements, models, and implications* - bron: COMPUTER NETWORKS, 20090318 - Vol. 53, Issue 4, p. 501-514

Finkelstein en David Salesin.¹⁰ Ze beschreven een wiskundige techniek, de zogenaamde *wavelet transformation*, om foto-beelden te indexeren. De afbeelding wordt geanalyseerd, en bij het zoeken wordt dit vergeleken met de index in de databank. Beelden die overeenkomen met het voorbeeld worden op die manier opgehaald. Hoewel de wiskunde erachter gecompliceerd is, is het principe eenvoudig, zoals de tekening op hun website illustreert:

Figuur 6: The Wavelet Transform

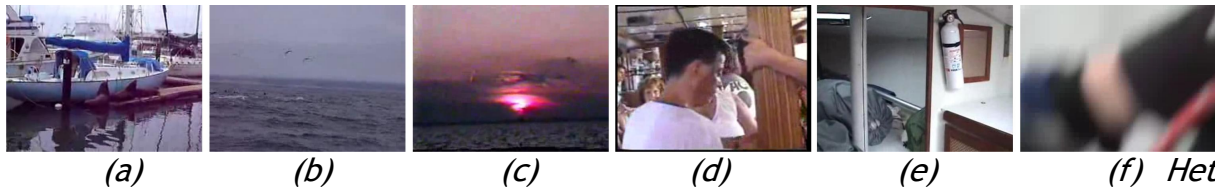


Shot Boundary Detection

Afzonderlijke beelden (frames) worden vergeleken tot de overeenkomst tussen volgend en voorafgaand beeld te groot is; op die plaats begint een nieuwe shot. Als de verdeling in shots gebeurd is, kan men hieruit representatieve beelden kiezen. Deze verzameling wordt een keyframe genoemd.

¹⁰ GRAIL, the Graphics and Imaging Laboratory of the University of Washington's Department of Computer Science and Engineering, <http://grail.cs.washington.edu/projects/query/>, geraadpleegd op 20091212

Figuur 7: Enkele monsters van keyframes geëxtraheerd uit een video met het trefwoord ‘zeilen’



Het toekennen van trefwoorden van dergelijk materiaal wordt nog verergerd door de complexiteit van de concepten (a, b), variërend uiterlijk (b, c), shots die niet direct visueel verband houden met zeilen (d, e), en met een lage kwaliteit van productie (f).¹¹

Om *Information Retrieval* te kunnen optimaliseren werd al heel wat onderzoek gedaan, zoals over het leggen van relaties tussen video's.

Hauptman et al. maken gebruik van een aanpak op basis van content-based copy retrieval (CBCR)¹² om 'near-duplicate video's' op te sporen met als doel de diversiteit in de zoekresultaten te bevorderen, door het verwijderen van overtollige zoekingen.

Near-duplicate web video's zijn dezelfde of ongeveer dezelfde video's als het origineel, maar in verschillende bestandsformaten, met gecodeerde parameters, fotometrische variaties (veranderingen in kleur en belichting), voorzien van bewerkingen (titel, logo en toegevoegde borders), van verschillende lengtes, en/of voorzien van bepaalde wijzigingen (frames toegevoegd / verwijderd).¹³

Een gebruiker zou een dergelijke video omschrijven als 'in wezen hetzelfde'. Liu et al. werken op dezelfde wijze, maar wegen ook tekst (text-based tools) om de relaties tussen de video's vast te stellen, en om zo te voorkomen dat verschillende interpretaties van dezelfde video aan de top van de resultatenlijst verschijnen.

In tegenstelling tot de vorige benaderingen, benutten Zink, Suh et al. visuele verbindingen en overlappingsen tussen video's ter verbetering van de kwaliteit en de homogeniteit van trefwoorden.

Siersdorfer et al. vermelden ook nog het 'FolkRank algoritme', dat wordt gebruikt om een rangschikking van trefwoorden te genereren voor een bepaalde gebruiker. Dit naar analogie van het principe *PageRank*, dat Google gebruikt om het web te indexeren.

Zij gebruiken een hybride benadering van de methode voor het wegen van termen in combinatie met CBCR voor het automatisch toekennen van annotaties.¹⁴

¹¹ Ulges A., Schulze C., Keysers D., Breuel T. - Content-based Video Tagging for Online Video Portals - bron:: Google Scholar, geraadpleegd op 20091515

¹² CBCR is gericht op het ophalen van alle gewijzigde versies of de vorige versies van een bepaalde video in een databank, in dit geval de databank van YouTube.

¹³ Xiao Wu, Alexander G. Hauptmann, Chong-Wah Ngo - Practical Elimination of Near-Duplicates from Web Video Search - bron: International Multimedia Conference, Proceedings of the 15th international conference on Multimedia, 2007 - p. 218-227

¹⁴ Siersdorfer S, San Pedro J, Sanderson M. - Automatic video tagging using content redundancy - bron: Annual ACM Conference on Research and Development in Information Retrieval archive Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009 - p. 395-402

Conclusie

Mede door de groei van de videomarkt en de economische impact die daarmee gepaard gaat, gebeurt het indexeren van videomateriaal voor het overgrote deel manueel: het toekennen van metadata door de gebruiker zelf. Deze bijdrage blijkt een relevante opbrengst te bieden en dit tegen een lage kostprijs. Huidige technieken voor indexering van beeldmateriaal leggen de focus op de analyse van de door de gebruiker toegekende tekst.

Aanvullend worden een hele reeks van technieken gebruikt voor het automatisch indexeren. Het grootste probleem lijkt de kostprijs voor het opzetten van een degelijk systeem te zijn.

Hoewel er veelbelovend onderzoek gaande is, zijn er nog geen bevredigende resultaten.

De reden dat er al zoveel onderzoek werd gedaan op delen van de databank van YouTube is wellicht het grote marktaandeel dat zij innemen. YouTube wordt daarbij geholpen door de mega-expertise van Google. Deze gigant heeft door de jaren heen al heel wat kennis opgedaan op vlak van *CBCR* (guess-the-google)¹⁵. YouTube plukt hier nu volop de vruchten van door gebruik te maken van een aantal van de technieken, die ik in deze paper uit de doeken heb gedaan.

Lijst met afbeeldingen

Figuur 1: Opbouw Website	4
Figuur 2: TopFrame	4
Figuur 3: zoekactie met als zoekwoord 'fótbolti's (IJslands)	7
Figuur 4: dezelfde zoekactie met als zoekwoord 'bóng đá' (Vietnamees)	7
Figuur 5: Video retrieval in YouTube	8
Figuur 6: The Wavelet Transform	9
Figuur 7: Enkele monsters van keyframes	10

¹⁵ <http://grant.robinson.name/projects/guess-the-google/>