Maskininlärningsbaserad indexering av digitaliserade museiartefakter

[Dnr:353-3849-2009]

Projektrapport

Lars Höglund, Johan Eklund och Kenneth Wilhelmsson

Enheten för biblioteks- och informationsvetenskap/
Institutionen för journalistik, medier och kommunikation*
Göteborgs universitet 2012

Tack

Tack till Etnografiska museet i Stockholm och speciellt till digitaliseringskoordinator Magnus Johansson som intresserat diskuterat projektet och bistått med testdata.

Detta projekt är finansierat av Riksantikvarieämbetet.

Lars Höglund Johan Eklund och Kenneth Wilhelmsson

Bakgrundsdata om projektet

Projektnamn: Maskininlärningsbaserad indexering av digitaliserade museiartefakter

Institution: Biblioteks- och informationsvetenskap/JMG Göteborgs universitet*)
Adress: Biblioteks- och informationsvetenskap
Institutionen för journalistik, medier och kommunikation
Box 710
Göteborgs universitet
405 30 Göteborg

Tel: 031-7864930
Mobil: 0708-773493

Medverkande: Lars Höglund, Göteborgs universitet, Biblioteks- och informationsvetenskap
Johan Eklund f.1970, Adjunkt, Högskolan i Borås, Biblioteks- och informationsvetenskap, Bibliotekshögskolan

Kenneth Wilhelmsson, f.1976, Fil.dr. Göteborgs universitet, Lingvistik

Projektledare: Lars Höglund, professor, f. 1946, Institutionen för journalistik, medier och kommunikation, Göteborgs universitet, Box 710, 405 30 Göteborg. *)

E-post: lars.hoglund@lis.gu.se

*) Den tidigare Enheten för biblioteks- och informationsvetenskap har 2011-07-01 förts till Institutionen för journalistik, medier och kommunikation vid Göteborgs universitet.

Innehåll

1 Sammanfattande resultat, bakgrund, syfte och material

Sammanfattande resultat

Projektet har genomfört försök med maskinbaserad analys och maskininlärning för automatisk indexering och analys av bilder som stöd för registrering av föremål i museibestånd. Resultaten visar att detta är möjligt för avgränsade delmängder i kombination med maskininlärning som stöd för, men inte som ersättning för, manuell analys. Projektet har också funnit behov av utveckling av ett användargränssnitt för både text och bildsökning och utvecklat en prototyplösning för detta, vilket finns dokumenterat i denna rapport och i ett separat appendix till rapporten. Materialet utgör grundunderlag för implementeringar som innebär utökade sökmöjligheter, effektivare registrering samt ett användarvänligt gränssnitt. Arbetet ligger i framkant av forskningsområdets resultat och etablerade metoder och kombinerar statististiska, lingvistiska och datavetenskapliga metoder.

Bakgrund och syfte

Projektets syfte är att i samråd med Statens museer för världskultur (SMVK), i synnerhet Etnografiska museet, studera metoder för maskinbaserad analys samt mönsterbaserad inlärning i samlingar av digitala bilder föreställande föremål i museernas bestånd. Målet med detta arbete är att belysa hur maskinell analys kan bidra till effektivare registrering av samt nya ingångar försökning på museiartefakter.

En stor andel av det digitaliserade material som rör museernas föremål utgörs av bilder. Det traditionella tillvägagångssättet för indexering av digitala bilder har varit att manuellt tilldela dessa standardiserade beskrivande poster bestående av egenskapsvärden, ämnesord, klassifikationskoder och annotationer. Denna process medger en hög noggrannhet i beskrivningen men innebär också en rad problem:

- Det manuella registreringsarbetet är ofta mycket tidskrävande.
- På grund av utsträckning över tid samt att flera personer eventuellt är inblandade i processen är risken stor att inkonsistenta beskrivningar uppstår i databasposterna och det digitaliserade materialet håller en ojämn kvalitet.
- I registreringsprocessen görs ett på förhand definierat urval bland de olika egenskaper hos föremålen som potentiellt skulle kunna användas i beskrivningen. Detta medför i sin tur att potentiellt relevant information riskerar att inte registreras och bli sökbar.

Projektet har särskilt studerat möjligheterna att maskinellt analysera bilder föreställande föremål i museisamlingar samt tillämpa maskininlärning för automatisk igenkänning av aspekter hos de avbildade föremålen, såsom föremålskategori, geografiskt ursprung, material samt tillverkningsmetod samt vidare beskrivit en prototyp för webb-baserad sökning av bild och text med förslag till användargränssnitt, där texter och sökfrågor representeras med s k vektorrymdsmodeller.

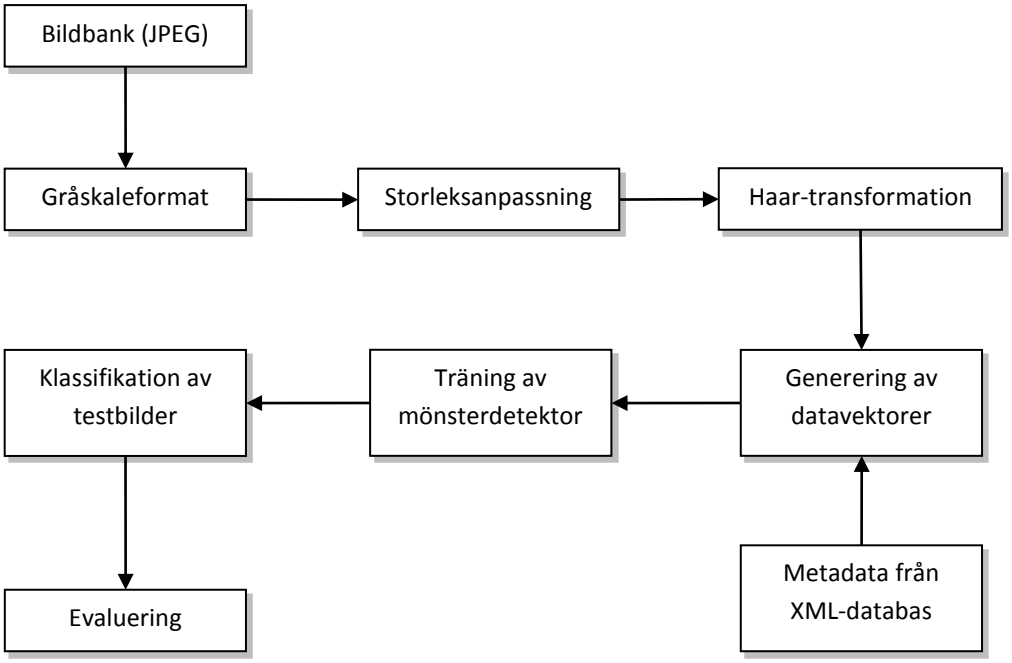
Material

Detta projekt har utförts med bildmaterial samt metadata tillhörande Etnografiska museets föremålsdatabas Carlotta. Samtliga bilder använda för mönsterinlärning har varit i formatet JPEG och används vanligen för visning av föremålen på Etnografiska museets webbplats.

2 Automatisk bildklassifikation

2.1 Tillvägagångssätt

För att möjliggöra maskinell analys av bildmaterialet samt träning för efterföljande klassifikation av föremålskategori, geografiskt ursprung, tillverkningsmaterial samt tillverkningsmetod har bildmaterialet först bearbetats i flera steg varefter varje bild har tilldelats en maskinläslig representation, kallad vektor, bestående av en lista med numeriska mätvärden – där varje mätvärde motsvarar en egenskap i bilden. En skiss över arbetsflödet visas i figur 1. Vid sidan om den maskinella analysen har metadata från databasen Carlotta överförts till XML-format och sedan använts för att tilldela kategorietiketter till bildernas representationsvektorer. Etiketter och bilddata har sedan använts för att träna en bildklassifikator (kallad mönsterdetektor i figuren) som bygger på en maskininlärningsalgoritm som kallas *supportvektormaskin* (SVM, se nedan).



Figur 1. Flödesdiagram över arbetsgången för automatisk bildklassifikation.

2.1.1 Vågelementtransformation

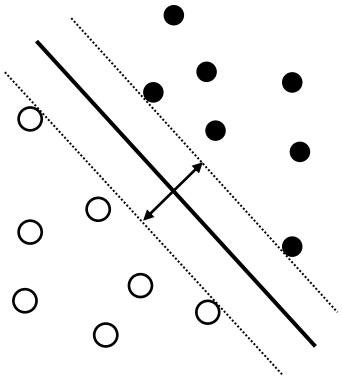
Metoder för detektering av mönster i bilder bör vara konstruerade så att de inte i någon påtaglig utsträckning påverkas detaljer som kan relateras till det enskilda fotograferingstillfället snarare än det avbildade föremålet, såsom ljus- och färgförhållanden samt föremålets exakta placering i bilden. De framträdande mönster som finns hos föremålen bör därför registreras utifrån relativa färgskillnader, snarare än specifika färgvärden, i bilden. För detta ändamål tillämpas inom bildanalys ofta en teknik som kallas *vågelementtransformation* (eng. *wavelet transform*). Denna teknik går i stora drag ut på att registrera skillnader mellan närliggande datapunkter i bilder på olika skal-/zoomnivåer, vilket gör det möjligt att fånga både makro- och mikrostrukturer i bilden. Den specifika teknik för

vågelementtransformation som har använts i detta projekt kallas *Haar-transformation*, vilket är ett vanligt val för bildanalys. Vågelementtransformation lämpar sig bäst för analys av mönster och strukturer i bilden, snarare än specifika konturer eller färgprofiler i bilden. Ett exempel på de första analysstegen i Haar-transformation ges i Appendix D.

I denna studie gjordes även preliminära tester att klassificera bilderna utifrån deras *färghistogram* (statistisk sammanställning av bildens fördelning av färgfrekvenser) men dessa medförde ingen hög klassifikationsprestanda och tillämpades inte för de resultat som presenteras nedan.

2.1.2 Supportvektormaskiner

Supportvektormaskiner (eng. support vector machines) är benämningen på en familj av metoder för problem inom maskinell klassifikation och regressionsanalys (generering av parameterbaserade modeller som kan användas för att beskriva datamängder och estimera (okända) värden i dessa datamängder). Supportvektormaskinen tillämpar det inom maskininlärning vanliga tillvägagångssättet för representation av objekt som *vektorer* (sekvenser av numeriska värden) samt avancerade geometriska metoder (för ett enkelt exempel, se skiss i figur 2) för att maximera maskinens klassifikationsprestanda. Den betraktas idag som en av de mest effektiva metoderna för maskinell klassifikation.



Figur 2. Supportvektormaskinen bygger på en inlärningsfas där en modell för geometrisk separation av kategorierna lärs in från träningsexempel (cirklar i figuren). Ett nytt objekt klassificeras utifrån vilken sida om det separerande planet (heldragen linje i figuren) objektets representationsvektor hamnar. I figuren ser vi två kategorier representerade av ifyllda respektive icke ifyllda cirklar.

Ett av de tillvägagångssätt som supportvektormaskiner använder för att maximalt separera kategorier i träningsdata är att tillämpa icke-linjär separationsmodeller (exemplet i figur 1 tillämpar linjär separation). Den tekniska termen för den mekanism som påverkar den geometriska formen för separationsmodellen är *kärna* (eng. kernel). I detta projekt har en s k polynomkärna använts, vilket innebär att vi har tillämpat icke-linjär separation för optimal klassifikationsprestanda.

2.2 Resultat

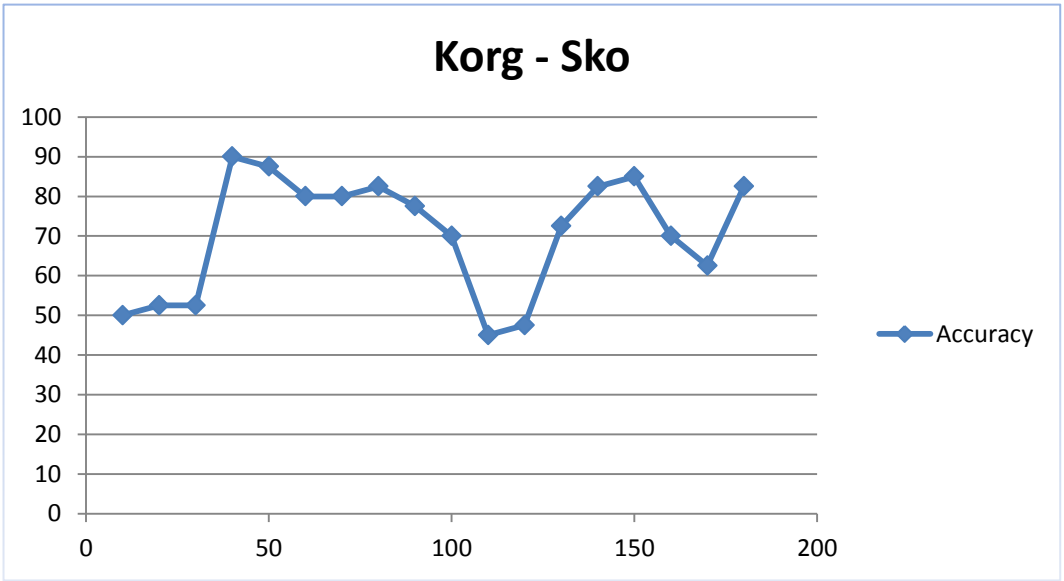
Som utvärderingsmått för denna delstudie används *accuracy*, definierat som andelen korrekt klassificerade bilder, dvs

$$\text{accuracy} = \frac{\text{antal korrekt klassificerade bilder}}{\text{antal klassificerade bilder}} \text{ x } 100$$

För träning av bildklassifikatorn har bilder från respektive kategori slumpvis valts ut så att träningsmängden består av lika många bilder från respektive kategori. Den inducerade klassifikatorn har sedan utvärderats i en separat testmängd bestående av 20 bilder från respektive kategori. Utvärderingen har skett genom att klassifikatorn har tränats över olika stora träningsmängder och sedan evaluerats i testmängden. I resultatdiagrammen anges accuracy som beroende variabel av storleken på träningsmängden. Som klassifikator har genomgående en supportvektormaskin med polynomkärna av graden 2 använts. Vi presenterar nedan ett urval av resultat som erhölls i denna delstudie.

Uppgift 1: Korrekt igenkänning av och differentiering mellan kategorierna Korg respektive Sko.

Denna klassifikationsuppgift bestod i att för en given bild korrekt välja mellan kategorierna *Korg* respektive Sko. För denna delstudie används 711 bilder för kategorin *Korg* och 201 bilder för kategorin Sko. Utvärderingsresultatet för klassifikationen av testmängden av bilder visas i figur 3. Den maximala klassifikationseffektiviteten (90% accuracy) uppnåddes med en träningsmäng på 40 bilder. Genomsnittlig accuracy över samtliga storleksnivåer var 70,6% (medianvärde 75%). En observation som kan göras i diagrammet är att den högsta klassifikationsprestandan uppnåddes vid relativt små respektive relativt stora träningsmängder. En möjlig förklaring är att de mindre träningsmängderna är mer homogena vad avser de ingående föremålens struktur, medan stora träningsmängder bidrar till en rikare klassifikationsmodell som är anpassad för variationer i materialet.



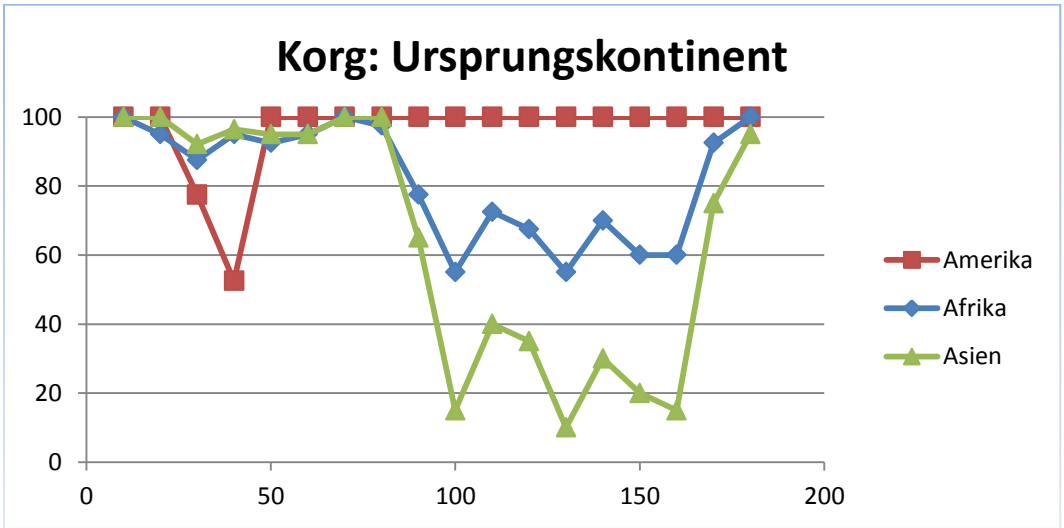
Figur 3. Accuracy för bildklassifikation i syfte att särskilja kategorierna Korg och Sko.

Uppgift 2: Identifikation av ursprungskontinent i föremålskategorin Korg

Denna klassifikationsuppgift bestod i att identifiera ursprungskontinent hos föremål i kategorin Korg. Syften med denna uppgiften var alltså att studera möjligheten att maskinellt identifiera geografiskt ursprung. Resultatet från utvärderingen visas i figur 4. Genomsnittlig accuracy för ursprungskontinenterna var som följer:

Amerika: 96,1% (median 100%)
Afrika: 81,8% (median 90%)
Asien: 65,5% (median 83,6%)

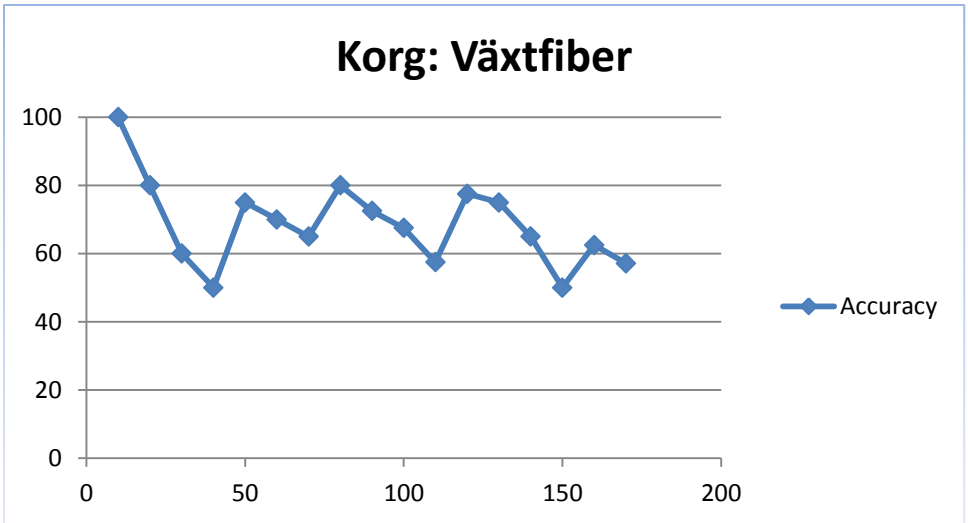
En tendens liknande den i föregående uppgift kan skönjas med avseende på klassifikationskorrekthet, dvs högst accuracy uppnås för relativt små respektive relativt stora träningsmängder, med en minskning för medelstora träningsmängder.



Figur 4. Accuracy för bildklassifikation i kategorin Korg med avseende på geografiskt ursprung (kontinent).

Uppgift 3: Identifikation av tillverkningsmaterial

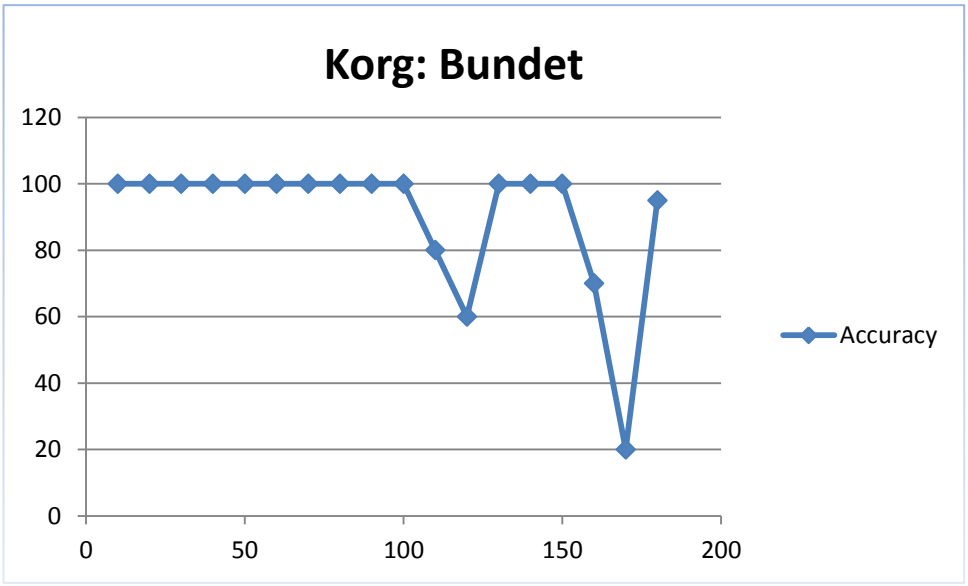
Denna uppgift bestod i att identifiera föremål i kategorin Korg tillverkade av materialet *växtfiber*. Resultatet från denna delundersökning redovisas i figur 5. Genomsnittlig accuracy för denna uppgift var 68,5% (median 67,5%).



Figur 5. Accuracy för klassifikation av föremål i kategorin Korg tillverkade av växtfiber.

Uppgift 4: Identifikation av tillverkningsmetod

Denna uppgift bestod i att identifiera föremål i kategorin Korg tillverkade av *bundet* material. Resultatet från denna delundersökning redovisas i figur 5. Genomsnittlig accuracy för denna uppgift var 90,3% (median 100%).



Figur 6. Accuracy för klassifikation av föremål i kategorin Korg tillverkat i bundet material.

2.3 Analys av semantiska relationer mellan deskriptorer

Föremålsposterna i databasen Carlotta är försedda med en rik uppsättning fält för beskrivning av föremålens egenskaper. De fältvärden som tilldelas posterna, särskilt de fältvärden som utgörs av enstaka ord från en begränsad deskriptorvokabulär, kan med ett sammanfattande ord

kallas deskriptorer. En lingvistisk teori som går under namnet *distributionshypotesen* uttrycker att ord som används i liknande textuella sammanhang också är semantiskt relaterade. Denna hypotes ligger till grund för en forskningsinriktning inom informationsvetenskap som syftar till att identifiera, strukturera, visualisera och tillämpa semantiska relationer för exempelvis förbättrade söktjänster och automatisk innehållsgruppering av dokument.

Fälten för tillverkningsmaterial respektive tillverkningsmetod i Carlotta innehåller i flertalet fall flera deskriptorer per föremålspost, vilket har gjort det möjligt att analysera och visualisera semantiska relationer mellan dessa beskrivningsord. Som mått på styrkan hos den semantiska relationen har vi konstruerat ett avståndsmått baserat på idéer från området *informationsteori*.

2.3.1 Informationsteori och semantiska relationer

Inom den grundläggande informationsteorin studeras en kommunikationsmodell som bygger på att informationspaket i form av meddelanden (eller signaler) skickas från en mottagare till en sändare. Då överföringskapaciteten hos kommunikationsmediet (exempelvis en radiosignal eller digital överföring i nätverk) är begränsad är det av intresse att studera hur informationsmängden hos översända meddelanden kan kvantifieras samt hur meddelandet kan kodas för att maximalt utnyttja överförda informationsenheter. Ett mått som formulerats inom ramen för informationsteori kallas *entropi* och kan informellt förstås som en kvantifiering av kapaciteten hos ett överföringsmedium, baserat på mängden av möjliga signaler som kan överföras via detta medium. Entropin för en enskild informationsenhet i ett meddelande kallas även *självinformation* (eng. self-information) och uttrycker på sannolikhetsmässig grund mängden information i informationsenheten. I denna studie har ett entropibaserat mått på semantiskt avstånd mellan termer använts för att kartlägga termernas inbördes relationer. I appendix C ges en formell presentation av detta avståndsmått.

2.3.2 Multidimensionell skalning

Multidimensionell skalning (eng. *multidimensional scaling*, MDS) är en teknik för approximering av multidimensionella dataenheter i ett lågdimensionellt rum, exempelvis ett plan. Med ”multidimensionell dataenhet” avses här en representation av ett objekt (exempelvis ett dokument eller ett föremål) bestående av en lista av numeriska mätvärden, där varje mätvärde korresponderar mot en specifik egenskap hos objektet. MDS är därför särskilt användbart för visualisering av relationer mellan dataenheter i diagrammatisk form. Det relativa avståndet mellan två punkter i diagrammet återspeglar likheten mellan de två objekt som dessa punkter representerar – ju större avstånd mellan punkterna i diagrammet, desto mindre likhet mellan de korresponderande objekten (och vice versa).

Ur tekniskt perspektiv utförs analysen genom att dataenheterna initialt ges en slumpmässig position i diagrammet, varefter MDS-algoritmen iterativt beräknar nya positioner för dataenheterna i syfte att successivt åstadkomma en representation med så litet approximationsfel som möjligt. MDS har i denna studie använts för att visualisera relationer mellan föremålsdeskriptorer, baserat på de informationsteoretiska relationsmått som beskrivits ovan samt i appendix A.1-4.

2.3.3 Hierarkisk klusteranalys

Hierarkisk klusteranalys syftar till att generera en partitionering (uppdelning) av objekt i mindre grupper, benämnda kluster, som i sin tur delas in i grupper på högre nivå så att en hierarkisk struktur av kluster uppstår. En vanlig visualiseringsform för dessa strukturer är träddiagram, även kallade *dendrogram*. Till skillnad från MDS, där relationer visualiseras genom relativa avstånd, leder klusteranalys till en skarp indelning av objekt i grupper utan bevarande av gradvisa förhållanden mellan objekt eller kluster. Klusteranalys är således användbart då man önskar erhålla skarpt avgränsade grupper av objekt med hög inbördes likhet. Annorlunda uttryckt är klusteranalys en teknik för automatisk klassifikation av objekt som, till skillnad från övervakade metoder (se avsnittet om supportvektormaskiner ovan), sker utan ett förberedande träningssteg.

Hierarkisk klusteranalys har i denna studie använts som alternativ till MDS för att generera och visualisera strukturer av föremålsdeskriptorer. Ett exempel på dendrogram producerade med denna metod presenteras i appendix B.

2.4 Sammanfattande analys

Utifrån, de till omfattningen relativt begränsade, resultaten från denna studie är intrycket att maskinell klassifikation kan vara ett användbart redskap vid arbetet med att registrera föremål. För några av klassifikationsuppgifterna uppnåddes en god säkerhet vid utvärderingen, men ett generellt problem är att avgränsa en träningsmängd av bilder som passar för en specifik registreringsuppgift. Som framgår av resultatpresentationen var det en stor variation i accuracy över antalet bilder i träningsmängden. Eftersom bildsamlingen i detta fall var relativt heterogen, dvs föremål i samma föremålskategori uppvisar sinsemellan stora olikheter, är det en rekommenderad strategi att låta den maskinella klassifikationen utföras i avgränsade och smala kategorier. Det bör dock påpekas att maskinell analys och klassifikation inte bör användas för att ersätta, utan för att komplettera den manuella processen genom att systemet vid registrering av nya föremål ger förslag på basis av evidens inhämtad från bilder med fullständiga poster.

Semantisk analys av bilddeskriptorer är ett forskningsområde med solid teoretisk bakgrund och flera viktiga tillämpningsområden i såväl registrerings- som sökprocessen. Genom att kartlägga deskriptorernas inbördes relationer kan registratören förses med uppgifter om vilka deskriptorer som tenderar att användas i liknande sammanhang och detta blir i sin tur ett stöd för en mer konsistent registrering. På motsvarande sätt kan en visualisering av deskriptorernas relationer vara en hjälp vid sökning på föremål då användaren ges möjlighet att navigera mellan termerna och bilda sig en uppfattning om det bakomliggande datamaterialet. Vidare kan klusteranalys tillämpas av systemet för sökning genom utifrån angivna söktermer föreslå ytterligare (kompletterande) söktermer genom extraktion ur klustren runt söktermerna. Det krävs dock ytterligare studier med det befintliga materialet för att få en uppfattning om i vilken utsträckning en sådan strategi medför förbättringar av användarnas sökresultat. För detta ändamål avser vi skapa en prototyp för ett sökgränssnitt där såväl automatisk klassifikation av bilder som statistiskt genererade ordlistor används för utöka sökmöjligheterna. I en uppföljningsfas till denna initiala studie ämnar vi därför mer ingående undersöka användbarheten för registrering och sökning hos automatiskt genererade ordlistor enligt ovan beskrivna metod.

3 Design av prototyp för sökning

3.1 Bakgrund – projektets innehåll och tidigare projektdel

Detta avsnitt behandlar förlängningen som genomfördes under 2011 och rör skapande av ett webb-gränssnitt för att praktiskt söka efter museiföremål med hjälp av exempelbilder eller med text. Det innebär först och främst implementation av sökalgoritmer för vektorrepresentationer av bilder av föremål (*cosinus*-likhet samt *euklidiskt avstånd*) för sökning med av användaren valda bilder på föremål från Etnografiska museet i Stockholm.[1] (I detta fall föreställer de aktuella bildposterna uteslutande skor och korgar från samlingarna.) De algoritmer som här används för bildsökning med vektorer hade en tidig praktiskt tillämpning i *SMART*-projektet (Salton, 1971) vid Cornell University på 1960-talet, där texter och sökfrågor representerades med s k vektorrymdsmodeller.

Från den föregående delen av detta projekt har vektorrepresentationer tagits fram för de drygt 900 föremålen.[2] Varje bild representeras av en vektor med 3072 elementvärden. När en sökning med hjälp av bilder gjorts av användaren är det vid uppvisande av resultatet också möjligt att markera relevanta svarsbilder och därefter återupprepa sökfrågan med denna extra information (återkoppling/feedback, med *Rocchio*-algoritmen, se t.ex. Manning et al, 2008). Detta kan leda till en förbättrad sökfråga och därmed ett bättre resultat, vilket därefter visas.

Till varje föremålspost finns även beskrivande textdata som visas upp i gränssnittet och som kan användas för textsökning, se tabellen nedan. De olika beskrivningsaspekterna kan lämpligen ses som ”kolumnrubriker” i en tabell där varje rad innehåller en föremålspost. Den befintliga ifyllda informationen motsvarar då emellertid en mycket glest ifylld tabell. Textdata är alltså välstrukturerad i olika beskrivningsfält men få fält är ifyllda för alla poster, se Tabell 1.

Det är möjligt att söka efter textinformation i samtliga fält (”kolumner”) där det finns information (se nedanstående tabell). Samtliga poster har alltså *id* och *sakord* ifyllda medan t.ex. *Utlånat* bara är ifyllt för tre föremål. I databasen som helhet förekommer många fler fält varav inga aktuella poster hade ifyllda värden i. Kolumnnamnen redovisas i befintligt skiftläge.

[1] Dessa föremål är hämtade från Carlotta-databasen.

[2] Se den tidigare beskrivningen av framställandet av dessa vektorer och utvärderingen, i projektrapporten för föregående del av detta projekt.

Kolumn	Antal	Kolumn	Antal	Kolumn	Antal	Kolumn	Antal
Sakord	911	**Höjd**	647	***Förvärvsomständigheter (i fält)***	182	***Lokalt namn***	71
id	911	**Teknik (tillverkning)**	621	***Ort (ursprung)***	182	***insektssanerat***	62
Datum (dokumentation)	889	**OWC (ursprung)**	564	***Datum (förvärv till museet)***	159	***Förvärvsomständigheter (till museet)***	54
Personnamn (dokumentation)	885	**Diameter**	414	***Inventarienummer, tidigare***	139	***Land (tillverkare)***	53
Ursprungsdatabas	865	**Datum (uppdatering)**	372	***Delar***	101	***Inventeringsstatus***	50
Inventarienummer	829	**Etnisk grupp (ursprung)**	366	***Specialbenämning***	98	***allmän kondition***	49
Beskrivning	827	**Personnamn (uppdaterare)**	344	***Tillstånd***	87	***Utställning (tidigare)***	48
Världsdel	815	**Bredd**	299	***gammal placering***	85	***OWC (brukare)***	45
Material (tillverkning)	765	**Personnamn (förvärvat från)**	293	***Anmärkning***	83	***Ort (brukare)***	44
Land (ursprung)	739	**Sakord engelska**	266	***Delobjektnummer***	82	***foto_dok_dig***	43
Materialkategori	734	**Region (ursprung)**	248	***Källor***	76	***Utställning, del av (tidigare)***	42
OCM	733	**Längd**	215	***Institutionsnamn (förvärvat från)***	73	***Datum (förvärv i fält)***	41

Kolumn	Antal	Kolumn	Antal	Kolumn	Antal	Kolumn	Antal
Utställning (ingår i)	40	***del***	12	***utfort_arbete***	4	***Djup***	1
skador	39	***Sakord kikongo***	12	***ovriga_anteckningar***	4	***Beskrivning, engelska***	1
Etnisk grupp (brukare)	35	***Etnisk grupp (tillverkare)***	12	***Sakord japanska***	4	***Personnamn (tillverkare)***	1
forvaring	34	***Sakord ainu***	12	***Sakord kibembe***	4	***Påskrift***	1
Utställning, del av (ingår i)	33	***Sakord kisundi***	12	***Utlånat***	3	***Sakord chokwe***	1
Land (brukare)	26	***utstellningsmontering***	10	***tidsatgang***	3	***Sakord tigrinja***	1
Tidpunkt (tillverkning)	22	***N***	10	***Omkrets***	3		
Region (tillverkare)	21	***OWC (tillverkare)***	10	***Publicerad text***	3		
Region (brukare)	19	***foto_dok_dia***	7	***Tidpunkt (användning)***	3		
Sakord, vidare term	16	***Funktion***	6	***Vecka (Veckans föremål)***	2		
Pris	16	***utstellt***	5	***arbete_utfort_av***	2		
Ort (tillverkare)	13	***Tidigare sakord***	5	***Motiv***	2		

Tabell 1 Efter borttagande av poster där text eller bild saknas har de 911 olika föremålsposterna ifyllda värden enligt ovanstående tabell.

3.2 Sökaspekter: bild och text

Slutresultatet för denna projektdel är en webb-baserad implementation med användargränssnitt som tillåter sökning med hjälp av exempelbilder (*query-by-example*) och en möjlig påföljande återkopplingssökning där relevanta objekt i svarsmängden markeras av användaren. Programmet tillåter även textsökning mot all text som hör till föremålen eller ett speciellt valt fält.

Gränssnittet har gjorts i tre delar (tekniskt sett sk. *iframes*) för sökning (till vänster) se Figur 1 och 2, uppvisande av resultat och markering av poster (i mitten), och detaljinformation om en enskild markerad post (till höger). I högervyn i gränssnittet finns även hjälptexter i en flik.

Figur 6 Bildsökning i gränssnittet.

Figur 7 Textsökning i gränssnittet.

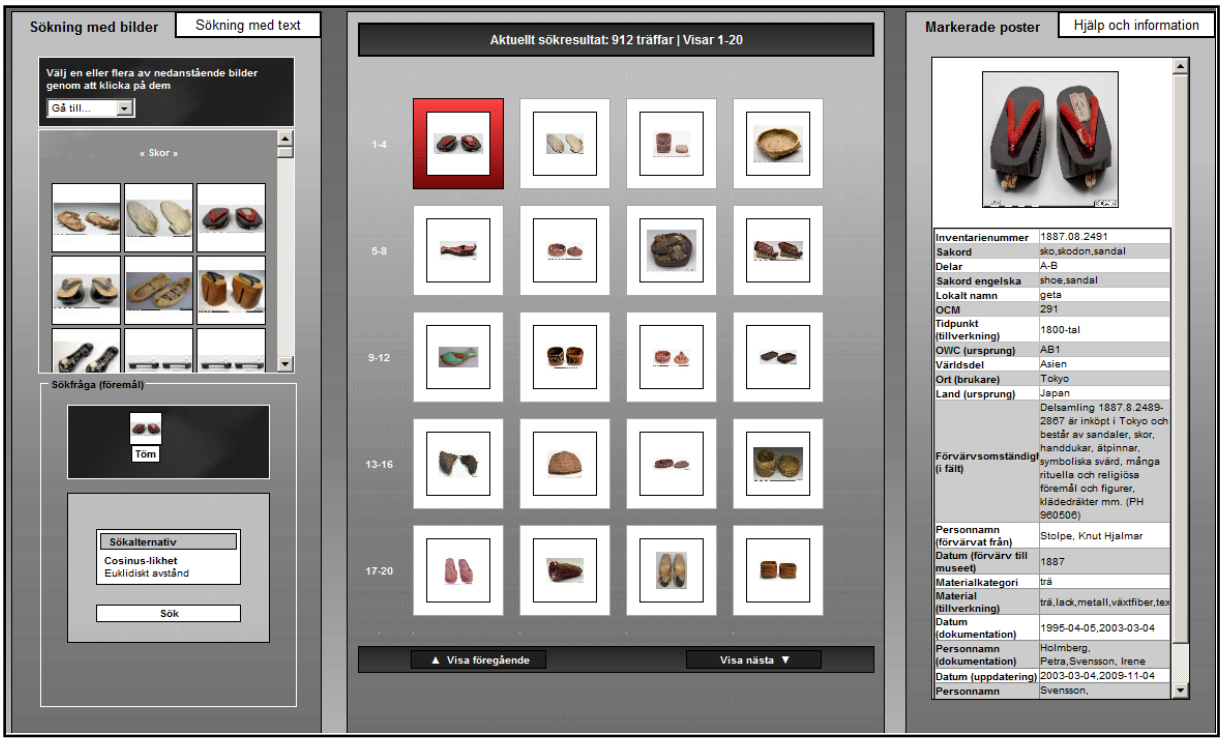
3.3 Den aktuella tekniken och miljön

Det som tagits fram är en prototyp som fungerar för Microsoft Internet Explorer under Windows (företrädesvis en sen version). Programmeringen är uteslutande gjord i JavaScript/Jscript (även förarbetet) och med (D)HTML och grafik för användargränssnittet.

Hela databasen av text- och vektordata är helt omgjord till JavaScript-format och programmet är helt och hållet gjort för att köras på klienten. Detta innebär en viss accentuerad inladdningstid men därefter i allmänhet ett snabbt och responsivt användargränssnitt.

Programmeringen är funktionsbaserad med text- och vektordata i anpassade associativa arrayer, se vidare nedan. Sökalgoritmerna som här gjorts i JavaScript är återimplementationer från den föregående projektdelen då korrektheten även utvärderades numeriskt.

3.4 Beskrivning av det grafiska gränssnittet



Figur 8 Gränssnittet i bildsökningsläge.

I Figuren ovan visas gränssnittet tre delvyer; vänstervyn med bildsökning, mittvyn med uppvisande av sökresultatet och där ett föremål valts för återkopplingssökning och därmed rödmarkerats. I högervyn syns detalj-information om det aktuella markerade föremålet.

Sökning med bilder

När det gäller praktisk sökning är denna mycket snabb när data är inladdad. Bildsökning med cosinusmåttet eller euklidiskt avstånd innebär följande övergripande programprocedur:

1. Användaren har formulerat en sökfråga genom att ha valt en eller flera bilder.
2. En centroidvektor räknas fram för de olika bilder som används.
3. Centroidvektorn (dvs. sökfrågevektorn) jämförs med det angivna bildlikhetsmåttet (cosinus-likhet eller euklidiskt avstånd) mot samtliga övriga föremålsbilder.
4. Föremålen rangordnas efter bildlikhet.
5. Sökresultatet presenteras för användaren.

Bildsökning med återkoppling – Rocchio-algoritmen

I programmets bildsökningsdel ingår ett återkopplingssteg (feedback). Det innebär att användaren kan markera ett antal bilder i sökresultatet som relevanta och bra svar. Användaren väljer återkopplingssökning efter att bilder markerats och åstadkommer därmed en modifiering av sökfrågan. De bilder som markerats används för att beräkna om sökfrågan och ge ett nytt modifierat sökresultat.

Textsökning

Textsökningen sker mot en objektbaserad databas som är inladdad till klienten. Detta möjliggör mycket snabb sökning. Exempelvis kan en sökfråga tänkas samman av flera kriterier och den delmängd av poster som matchar ett visst villkor kan direkt visas.

Från ett gränssnittsperspektiv är frågan hur den goda strukturen kan användas trots att fälten är så glest ifyllda. Den konkreta svårigheten ligger i att en så stor andel av de olika fälten (”kolumnerna”) har ifyllda värden. Det innebär att användaren löper stor risk att vid sökning med en sådan kolumn fylla i ett värde som inte matchar några poster alls.

Lösningen består av att dels ge direkt feedback vid ifyllt värde, om hur många poster som träffas (eventuellt visas också dessa upp direkt) och dels i att använda en autokompletteringsfunktion som hjälper användaren att fylla i ett värde som kommer att ge träffar.

Textsökningen som presenteras innebär dels en generell sökfunktionalitet som söker i samtliga fält och dels en avancerad funktionalitet för att kombinera fältvärden.

Se appendix för de hjälptexter och instruktioner som återfinns för praktisk sökning i det grafiska gränssnittet.

3.5 Representation och omarbetning och av datamängderna

De drygt 900 föremålen ur Etnografiska muséets Carlotta-databas har för varje post tre beskrivningsaspekter, vilka blir detta projekts datamängd.

- En bild för varje föremål, som används för uppvisande i gränssnittet. Olika storlekar har tagits fram av praktiska skäl. Formatet är JPEG.
- En bildvektorrepresentation utgående från varje bild. Denna bildvektor har för varje föremål 3072 elementvärden och har tagits fram i den föregående delen av detta projekt. Det är bildvektorn som är den komponent som praktiskt används i bildsökningen (cosinuslikhet och euklidiskt avstånd) och i återkopplingssökning (Rocchio-algoritmen).
- En textmässig beskrivning av föremålet i många olika fält (se Tabell 1), som erhållits i ett XML-format men som omstrukturerats av praktiska skäl, se nedan.

Eftersom den aktuella webb-versionen bygger på klientfunktionalitet finns två möjliga svårigheter i det faktum att mycket data laddas till klienten och att det är klientens webbläsare, operativsystem och hårdvara som avgör faktisk prestanda.

De delar av programmet som innebär en belastning och tidsfördröjning är för närvarande framför allt bildvektorerna (i sin första JavaScript-form ca 13,2 MB, se appendix) och i viss mån textdata och bilder.

Se appendix för algoritmer för sökning och en beskrivning av den omarbetning av bildvektorer och textdata som gjordes för att skapa en webb-implementation.

3.6 Kvarstående arbete

Efter projektet: artikelpublicering samt ev. diskussioner med museer om förutsättningar för implementering.

4 Referenser

- Manning, Christopher D., Raghavan, Prabhakar & Schütze, Hinrich (2008). Introduction to information retrieval [Elektronisk resurs]. Cambridge: Cambridge University Press.

  http://www-nlp.stanford.edu/IR-book/

- Salton, Gerard (red.) (1971). *The SMART retrieval system: experiments in automatic document processing*. Englewood Cliffs, N.J.

Till denna rapport finns ett appendix med ytterligare material.