

Web-based Information Content and its Application to Concept-Based Video Retrieval

Alexander Haubold
Dept. of Computer Science
Columbia University
New York, NY 10027
ahaubold@cs.columbia.edu

Apostol (Paul) Natsev
IBM Thomas J. Watson Research Center
Hawthorne, NY 10532
natsev@us.ibm.com

ABSTRACT

Semantic similarity between words or phrases is frequently used to find matching correlations between search queries and documents when straightforward matching of terms fails. This is particularly important for searching in visual databases, where pictures or video clips have been automatically tagged with a small set of semantic concepts based on analysis and classification of the visual content. Here, the textual description of documents is very limited, and semantic similarity based on WordNet's cognitive synonym structure, along with information content derived from term frequencies, can help to bridge the gap between an arbitrary textual query and a limited vocabulary of visual concepts. This approach, termed concept-based retrieval, has received significant attention over the last few years, and its success is highly dependent on the quality of the similarity measure used to map textual query terms to visual concepts.

In this paper, we consider some issues of semantic similarity measures based on Information Content (IC), and propose a way to improve them. In particular, we note that most IC-based similarity measures are derived from a small and relatively outdated corpus (the Brown corpus), which does not adequately capture the usage pattern of many contemporary terms: for example, out of more than 150,000 WordNet terms, only about 36,000 are represented. This shortcoming reflects very negatively on the coverage of typical search query terms. We therefore suggest using alternative IC corpora that are larger and better aligned with the usage of modern vocabulary. We experimentally derive two such corpora using the WWW Google search engine, and show that they provide better coverage of vocabulary, while showing comparable frequencies for Brown corpus terms. Finally, we evaluate the two proposed IC corpora in the context of a concept-based video retrieval application using the TRECVID 2005, 2006, and 2007 datasets, and we show that they increase average precision results by up to 200%.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIVR '08, July 7-9, 2008, Niagara Falls, Canada.

Copyright 2004 ACM 1-58113-000-0/00/0004...\$5.00.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing - *dictionaries, indexing methods, linguistic processing*. I.2.6 [Artificial Intelligence]: Learning - *concept learning, parameter learning*.

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Information Content, Semantic Similarity, Brown corpus, WordNet, TRECVID, LSCOM.

1. INTRODUCTION

Leveraging semantic resources, such as WordNet [1], has proven as a valuable linguistic tool in determining semantic relationships between words, sentences, and larger bodies of text [26]. WordNet's rich vocabulary and complex structure of relationships can be used to determine semantic concepts for words as well as links to near and distant related terms. With the emergence of audio-visual material and the desire to perform information retrieval on such resources, WordNet has become an important tool used in query expansion [2]. More recently, WordNet has become a medium used to bridge the gap between semantic and visual domains, as its concept classes are linked to visual counterparts [3].

Determining the semantic similarity between words or phrases is one of the most powerful features of WordNet, and it is frequently used for query expansion purposes. Many measures for semantic similarity have been introduced, the most popular of which use some combination of WordNet's hierarchy and a measure of information content to generate a numerical score of likeness. Such measures include Wu-Palmer [4], Resnick [5], Jiang-Conrath [6], Lin [7], Lesk [8], and Leacock-Chodorow [9], among others. Approaches by Resnick, Jiang-Conrath, and Lin rely heavily on the notion of information content, a numerical value which denotes a term's saliency, or importance.

A closer analysis of source of information content used with WordNet reveals that a measure of information content exists only for approx. 25% of all WordNet terms. Generally, a non-existent value eliminates a term's semantic relatedness to any other term. In this paper, we explore the reason behind this deficiency, outline the potential negative impact, provide alternatives, and evaluate

them on TRECVID search tasks [10, 11, 12]. For the work presented here, we have used WordNet version 2.1.

2. BACKGROUND

The WordNet lexical database models a large portion of words and phrases from the English language in a lexical hierarchy of cognitive synonyms with several types of relationships, such as hypernyms and synonyms. It is possible to resolve semantic relationships between concepts, such as the relationship between a leaf and a plant (hypernym), but also between named entities and concepts, for example, World Trade Center is “an instance of” a skyscraper and it is a type of building (hypernym).

Each of the four parts of speech captured in WordNet (noun, verb, adjective, adverb) underlies a different hierarchy of relationships. Nouns, outnumbering the other parts of speech, have the most comprehensive hierarchy, but at the same time the most straightforward linkage between terms, sharing only one common root named “entity”. It is therefore possible to find a path between any two nouns, whether this path traverses the root node of “entity”, or some other common sub-node. For example, the conceptually closely related terms “leaf” and “stem” share the common parent of “plant organ”. Only when terms are conceptually so distant that they do not share a common sub-node, such as the abstract term “thought” and the physical term “water”, is the more distant root node “entity” the only link.

Unlike nouns, verbs are grouped into smaller disjoint hierarchies, each one featuring one common node. These hierarchies are very shallow compared to those of nouns, and there exist no connections between them. For instance, the verb “to walk” is grouped into a hierarchy with root node “to travel”, for which no path exists to the verb “to drink” with root node “to consume”.

2.1 WordNet Similarity Measures

The hierarchy of concepts makes it possible to generate numerical measures of semantic similarity. Several such measures have been devised and evaluated for their semantic intuition. Depending on the individual approach, the numerical value of relatedness is based on some combination of the lowest common subsumer (LCS), which is the closest node shared by both concepts, the path distance between two concepts through the LCS, the distance from the hierarchy’s root, and other measures describing concept context. The most popular measures of semantic similarity include Wu-Palmer [4], Resnik [5], Jiang-Conrath [6], Lin [7], Lesk [8], and Leacock-Chodorow [9]. This collection has been implemented by [13] in a freely accessible Perl package.

Wu and Palmer’s similarity measure [4] is based on a depth ratio between two concepts ($c1$ and $c2$) and their lowest common subsumer. The ratio guarantees a value in the range of [0,1]:

$$sim_{wup}(c1, c2) = \frac{2 * depth(LCS(c1, c2))}{depth(c1) + depth(c2)} \quad (1)$$

where $depth(c)$ is defined as the distance between the concept c and the root node.

Several measures use the information content (IC) of terms to give more weight to specific and more informative terms, as measured by the term frequency. Intuitively, the information content of a term reflects its importance, or specificity, and is defined as:

$$ic(c) = \log\left(\frac{1}{prob(c)}\right) \quad (2)$$

where $prob(c)$ is the probability of a term c , which can be estimated from the term’s frequency in a large corpus.

Resnik [5] defines the similarity measure between two terms as the information content of the lowest common-subsumer between them:

$$sim_{res}(c1, c2) = ic(LCS(c1, c2)) \quad (3)$$

The range of values for this measure is [0, $\log(N)$], where N is the size of the IC corpus.

Jiang-Conrath [6] also uses information content in its similarity measure, but relates the IC value of the lowest common subsumer to the IC values of the individual concepts. The final value is inverted to guarantee a range of [0,1]:

$$sim_{jcn}(c1, c2) = \frac{1}{ic(c1) + ic(c2) - ic(l)} \quad (4)$$

$$ic(l) = 2 * ic(LCS(c1, c2))$$

Lin [7] incorporates information content into Wu-Palmer’s similarity by replacing the original measure of depth. This measure also guarantees a range of similarity values between [0,1], because the information content of the lowest common subsumer in the nominator of the equation is by definition lower than the IC values of the concepts in the denominator:

$$sim_{lin}(c1, c2) = \frac{2 * ic(LCS(c1, c2))}{ic(c1) + ic(c2)} \quad (5)$$

Leacock-Chodorow [9] base their similarity measure on shortest path length between two concepts, normalized by the overall depth of the taxonomy. The pre-final measure is guaranteed to be in the [0,1] range, but after the log computation, the measure exists in a space from [0,x], where $x = \log(2 * taxonomyDepth) > 1$:

$$sim_{lch} = -\log\left(\frac{shortestPath(c1, c2)}{2 * taxonomyDepth}\right) \quad (6)$$

The Lesk measure introduced by Banerjee and Pedersen [14] is based on a measure originally developed by Lesk [8] for sense disambiguation using dictionary definitions. The approach finds coinciding terms between the definitions of two concepts; the higher the “gloss overlap”, the more related the two concepts are. The numerical value for this measure can range anywhere between [1,x], where $x > 1000$.

2.2 Similarity Measure Evaluation

The collection of similarity measures has been evaluated to various standards and measures in several prior experiments. In [15], the authors evaluate semantic similarity by comparing results from automatic measures to human judgment. The authors from [16] have evaluated the similarity measures by applying

Table 1: Semantic similarity performance results from experiments published in four different research papers. Measures towards the top perform best. Large variation in rankings makes it impossible to select an overall best measure, but measures by Jian Conrath and Lin are consistently among the top-ranked.

| Seco [15] | Budanitsky [16] | Pucher [17] | Pedersen [18] |
|-------------|-------------------|-------------------------------------|----------------------|
| LCH (0.82) | Best: JCN | Best | Vector (0.76) |
| JCN (0.81) | Mid-way: LIN, LCH | measures: | LIN (0.69) |
| LIN (0.8) | Worst: RES, HSO | JCN, Path for nouns, LESK for verbs | JCN (0.55) |
| RES (0.77) | | | RES (0.55) |
| WUP (0.74) | | | Shortest Path (0.48) |
| LSA (0.72) | | | LCH (0.47) |
| HSO (0.68) | | | |
| LESK (0.37) | | | |

them for detection and correction of real-word spelling errors. In [17], the author evaluates similarity measures for prediction of words in conversational speech to improve speech recognition. In the medical domain, [18] has applied semantic similarity to a corpus of clinical terminology and patient records. A hierarchy similar to WordNet, but by a factor of 3 in size, was constructed for medical concepts. The authors evaluated performance of concept similarity measures by comparing them to human judgment. Results from these four evaluations have been included in Table 1 in order of performance.

It is evident that a similarity measure’s performance is highly dependent on the application, and results vary significantly. For example, while Lesk performs worst for comparison to human judgment in [15], it performs best for verb prediction in conversational speech [17]. Similarly, Leacock-Chodorow performs best for [15], and it performs worst in the application to the medical domain [18]. While there exists no universally applicable ground truth for any of the measures, measures such as Jiang-Conrath and Lin tend to perform average or above average. Along with Resnik, the computation behind these two measures depends highly on information content. However, due to the significant shortcomings of IC, we believe that the performance of these similarity measures could be improved with a more relevant IC database.

2.3 Information Content

Information content of a concept measures its ability to convey specific information. A highly recurring concept, for example, conveys little information due to its ubiquitous use, and therefore it has a small IC value. A rare concept conveys much more information because it tends to be highly specialized and conveys more independent meaning. Its IC value is much higher.

Information content in the context of WordNet is drawn from the Brown University Standard Corpus of Present-Day American English (a.k.a. the Brown Corpus). This corpus refers to a collection of documents of widely varying genres collected in 1961 (see Table 2), which was updated in 1971 and 1979 to

Table 2: Distribution of text sources in the Brown Corpus from 1961 – 1979. While there exists a wide variety, the contemporary equivalent to this list would include additional sources, in particular related to technology.

| |
|---|
| Press: Reportage (<i>Political, Sports, Society, Spot News, Financial, Cultural</i>) |
| Press: Editorial (<i>Institutional Daily, Personal, Letters to the Editor</i>) |
| Press: Reviews (<i>Theatre, Books, Music, Dance</i>) |
| Religion (<i>Books, Periodicals, Tracts</i>) |
| Skill and Hobbies (<i>Books, Periodicals</i>) |
| Popular Lore (<i>Books, Periodicals</i>) |
| Belles-Lettres (<i>Books, Periodicals</i>) |
| Miscellaneous: US Government & House Organs (<i>Government Documents, Foundation Reports, College Catalog, Industry House organ</i>) |
| Learned (<i>Natural Sciences, Medicine, Mathematics, Social and Behavioral Sciences, Political Science, Law, Education, Humanities, Technology and Engineering</i>) |
| Fiction: General (<i>Novels, Short Stories</i>) |
| Fiction: Mystery and Detective Fiction (<i>Novels, Short Stories</i>) |
| Fiction: Adventure and Western (<i>Novels, Short Stories</i>) |
| Fiction: Romance and Love Story (<i>Novels, Short Stories</i>) |
| Humor (<i>Novels, Essays</i>) |

reflect newer literature. The collection of more than 1 million words was manually tagged with about 80 parts of speech, such as singular nouns, plural nouns, verbs, various forms of adverbs, adjectives, pronouns, etc.

Parts of speech tags from the Brown Corpus do not correspond to the hierarchical structure of concepts in WordNet; in fact, there exists no meaningful mapping between parts of speech and concepts. In a separate manual step, a subset of texts from the Brown Corpus was tagged with WordNet concepts. In this manner, a semantic concordance was generated, with the Brown Corpus texts as a textual corpus and WordNet as a lexicon [19].

The presently available list of WordNet concepts tagged in the Brown Corpus includes approximately 420,000 words (WordNet’s *cntlist*). While there exist more than 155,000 unique words and phrases in WordNet, only about 37,000 of those coexist in the Brown Corpus.

The significant gap of missing concepts in the Brown Corpus consequently results in empty information content for roughly 75% of words and phrases from WordNet. Even though a large number of these are rarely used due to their specialized subject, many popular words are also among the list, such as “soccer”, “hockey”, “cpu”, “fruitcake”, or “World Trade Center”. The lack of their occurrence in the Brown Corpus can be explained by the shift in vocabulary use over the past 30 years. For example, it is plausible that “soccer” was either not heavily reported on in the period during which the Brown Corpus was created, or it was still referred to as “football”, which shares several entries in the

corpus. A term like “World Trade Center” was omitted, because the WTC buildings were erected after the initial compilation of texts. Yet other terms, like “cpu”, were simply not yet common, and have only been introduced in the common use of language due to the emerging underlying technology.

An equivalent to the Brown Corpus of today’s vocabulary would experience a shift in the genres from Table 2 so as to encompass a wider and more relevant set of vocabulary. For instance, the category of “Learned”, which includes all natural and social sciences, would likely be split into separate categories with added topics, such as computer science and biomedical engineering.

The numerical measure of information content of a WordNet concept is derived from the probability with which the concept occurs in the text corpus. WordNet concepts are first assigned the frequency values at which they occurred in the text corpus; those without any occurrence are assigned a zero. In the subsequent indexing step, frequency values are percolated from the leaves of the concept hierarchy towards the root. A node’s final frequency value is thus the sum of frequencies of its children. Consequently, the root has the highest frequency value, while the leaves have the lowest. Information content is then defined as follows:

$$ic(c) = -\log\left(\frac{freq(c)}{freq(root)}\right) \quad (7)$$

The definition of IC therefore assigns a value of zero to the root, because it contains no meaningful information, while assigning the largest IC values to the leafs of the hierarchy.

We note that equation (7) is equivalent to equation (2):

$$ic(c) = \log\left(\frac{1}{prob(c)}\right) \quad (2)$$

$$prob(c) = \frac{freq(c)}{freq(root)}$$

$$ic(c) = \log(1) - \log(prob(c)) \quad (7)$$

$$ic(c) = -\log(prob(c))$$

2.4 Shortcomings of Similarity Measures based on IC

Similarity measures that are based on path length and depth of concept nodes always resolve a non-zero value for concepts from the same hierarchy. When there exists no link between two concepts, for example between two verbs in different hierarchies, non-zero similarity can still be computed when gloss overlaps are used. Similarity measures that rely on information content, however, can produce zero-values for even the most intuitive pairs, because the majority of WordNet concepts occur with a frequency of zero. The measures Jiang-Conrath and Lin are particularly sensitive to this effect. When one of two concepts occurs with a frequency of zero then its information content is infinity in the continuous domain, and both similarity measures return zero. When both concepts occur with a frequency of zero, then, depending on whether their lowest common subsumer also occurs with a frequency of zero, Jiang-Conrath either returns zero or infinity in the continuous domain. With the same assumptions,

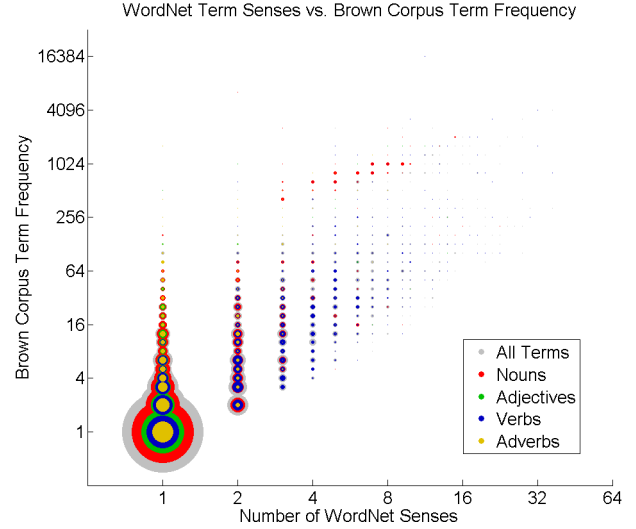


Figure 1: Most words and phrases in WordNet are mapped to only one concept. Also, most terms occur with a very low frequency in the Brown corpus. While sense disambiguation for WordNet concepts is generally favorable, only a small number of terms would benefit from this semantic step.

Lin’s measure would either return infinity or one. None of these results is truly indicative of the actual similarity, and therefore, in practical implementations, if a concept appears with a frequency of zero, then its information content is also zero. In case of an information content of zero, similarity measures return a zero.

3. Information Content from Alternative Corpora

Due to the large number of unaccounted concepts in the semantic concordance between the Brown corpus and WordNet concepts, the similarity measures of Jiang-Conrath and Lin do not reasonably reflect similarity for many concept pairs. The underlying database for information content needs to be revised, such that all WordNet concepts are associated to non-zero values of occurrence. Performing such a task manually is prohibitively expensive and therefore we present automatic methods of generating the frequency database.

The original concept frequency set based on the Brown Corpus was derived from manual concept tagging. The main difficulty of tagging words and phrases in a text with concepts lies in matching their meaning to the correct concept, both as part of speech (i.e. verb, noun, etc.) and as a sense (e.g. chemical compound “base” vs. military “base”). When a word has only one unambiguous meaning (e.g. “World Trade Center”, or “sofa”), mapping it to a concept is a straightforward task. However, when multiple senses exist, only the word’s context can help resolve the correct concept. When considering a much larger corpus for information content, it would be prudent to perform manual or semi-automatic mapping of words to concepts.

Analysis on the WordNet lexical database and the concept frequency set shows that most words and phrases are mapped to

Table 3: Top 39 domains from which web documents were used to create a WordNet word and phrase frequency database.

| | | |
|--|----------------------------------|--|
| en.wikipedia.org (73,932) | www.yourdictionary.com (8,522) | davesgarden.com (3,185) |
| www.thefreedictionary.com (68,958) | www.imdb.com (7,820) | plants.usda.gov (3,171) |
| www.answers.com (22,551) | www.merriam-webster.com (7,555) | www.geocities.com (3,140) |
| www.amazon.com (22,193) | www.online.thesaurus.net (7,183) | www.freepatentsonline.com (2,884) |
| www.wordwebonline.com (17,946) | www.wordreference.com (6,252) | www.infoplease.com (2,849) |
| books.google.com (15,633) | encyclopedia.farlex.com (5,933) | www.dictionarydefinition.net (2,670) |
| dictionary.reference.com (12,434) | links.jstor.org (5,096) | www.myspace.com (2,656) |
| dict.die.net (11,042) | images.google.com (4,880) | www.flickr.com (2,534) |
| www.britannica.com (10,525) | www.ncbi.nlm.nih.gov (4,825) | www.wordnet-online.com (2,506) |
| www.elook.com (9,892) | www.crosswordsolver.org (4,316) | animaldiversity.ummz.umich.edu (2,499) |
| www.bartleby.com (9,394) | www.youtube.com (3,651) | news.google.com (2,469) |
| onlinedictionary.datasegment.com (9,242) | findarticles.com (3,403) | www.allwords.com (2,378) |
| cancerweb.ncl.ac.uk (8,604) | thesaurus.reference.com (3,311) | www.nlm.nih.gov (2,206) |

Table 4: Top- and bottom-ranked document frequencies based on a 1 million + sample set of web documents.

| | | | |
|-----------------|-----------------|----------------------------|-------------------------------|
| a (928,560) | 1 (601,885) | habitableness (3) | witches' broom (1) |
| in (902,187) | more (597,140) | maitre d' (3) | faggpting (1) |
| by (770,365) | be (590,466) | 's gravenhage (2) | philosophers' stone (1) |
| on (766,855) | other (586,855) | misspeka (2) | old wives' tale (1) |
| or (716,489) | at (583,129) | walter john de la mare (2) | witches' brew (1) |
| all (713,604) | an (574,730) | miniconju (2) | archimedes' principle (1) |
| new (633,461) | as (567,580) | davis' birthday (2) | yorktwon (1) |
| about (633,314) | are (562,785) | ikhanaton (2) | april fools' (1) |
| it (615,344) | page (549,403) | stevens' power law (1) | jefferson davis' birthday (1) |
| us (602,074) | not (544,547) | partitia (1) | achilles' heel (1) |

only one concept. Furthermore, we observe that most words occur with very low frequencies between 1 and 5. Figure 1 presents these observations in a two-dimensional histogram of terms mapped to sense count and frequency. The histogram shows mappings of individual parts of speech including nouns, adjectives, verbs, and adverbs, as well as all parts of speech combined. Clearly, the number of terms belonging to only one WordNet concept far outnumber the remaining ones. Our analysis suggests that for most words and phrases in WordNet, laborious tagging is unnecessary due to the relatively small number of terms requiring disambiguation. A fully automatic approach to computing information content using mere term frequency counts is therefore possible, while not perfect. We present two such approaches using widely accessible data from the WWW.

3.1 Sample of WWW Pages

We have collected a sample of 1,231,929 unique web pages containing 1,169,368,161 words and phrases from the WordNet dictionary. From this corpus, term frequencies are used to generate a new information content database. The sample of web pages was systematically selected from the top-ten Google search results for each WordNet concept. While this process is not supervised, as was the selection of the Brown corpus, it relies on the general perception that Google's top search results are popular and, to the most part, relevant.

We now describe the process used to collect the corpus of web documents. The underlying manual procedure is analogous to performing a standard web search for a term or phrase, e.g. "circus", using a popular search engine, e.g. Google. One-by-one, we archive the HTML source for each of the ten top matching web pages in the search result list.

To collect a large corpus of web documents, which is distributed over all WordNet concepts, we automate the above-described procedure. We first establish a list of WordNet words and phrases, each representing one search query generating ten search results (URLs of various web pages). The list is a concatenation of nouns, verbs, adjectives, and adverbs denoting WordNet concepts, extracted from WordNet's index files (*index.{noun,verb,adj,adv}*). WordNet contains 155,327 unique terms when considering part of speech, and 125,209 terms when disregarding part of speech. For example, the term "shelf" exists only as a noun, whereas the term "go" can occur as a noun, verb, and adjective. Since there currently exists no method of searching for specific parts of speech using mainstream search engines like Google, we disregard part of speech and generate one dictionary of unique words and phrases.

An automated script performs the search for each word and phrase in this list. Search queries for phrases, which contain more than one word, are double-quoted to guarantee phrase-level matches,

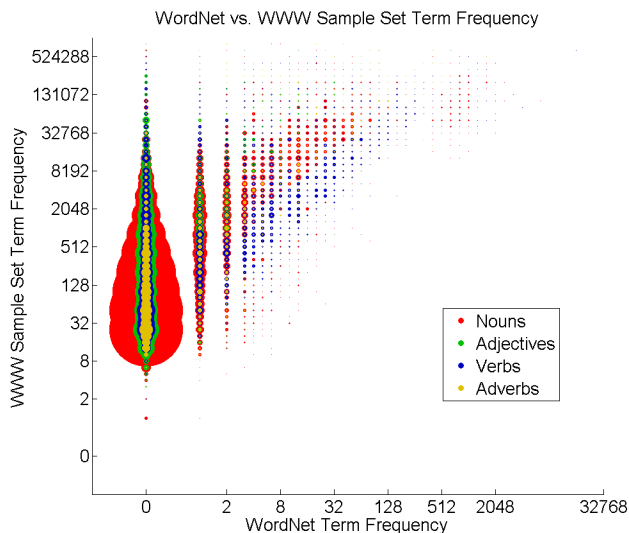


Figure 2: Comparison of term frequencies between WordNet based on the Brown corpus and a large sample set of web pages. Most WordNet terms are not accounted for in the Brown corpus, which is evident from the large distribution in WordNet Term Frequency bin zero.

i.e. “World Trade Center” instead of *World Trade Center*. The top-ten matching web documents for each query are retrieved and stored. Duplicate web page matches occurring in search results from different queries are counted as only one result. Of 125,209 search queries submitted, 98.39% of all top-ten results were unique. Web domains with the most hits include Wikipedia and The Free Dictionary. Table 3 lists the top 39 domains in our corpus. With a total of 400,240 web documents, they account for one third of all web pages analyzed. Among them are predominantly dictionaries and other mainstream information sites and medical portals.

Raw data in the resulting corpus of 1,231,929 documents is stripped of all non-text elements, including HTML tags and binary data, e.g. from PDF documents. Also removed are titles and other elements that do not appear in a coherent body of written material. The remaining text is then filtered for words and phrases that occur in the WordNet database, and aggregated in a term frequency table. Table 4 lists top- and bottom-ranked frequency results. While we expect the lowest term frequency to be 10 (at least as many as there are top-ten web pages in the search results), the actual values of the lowest-ranking terms are in fact less due to the aforementioned text filter step.

We derive information content from this distribution in a process similar to that introduced by Pedersen and Patwardhan (*semCorFreq.pl*). In a top-down approach, frequency values are assigned to WordNet word senses, and are then percolated to the corresponding hypernyms. Leaf nodes therefore carry the actual frequency values, while hypernyms represent sums of frequencies of their children.

The approach differs from *semCorFreq.pl* in two details. First, since the frequency table represents words and phrases without

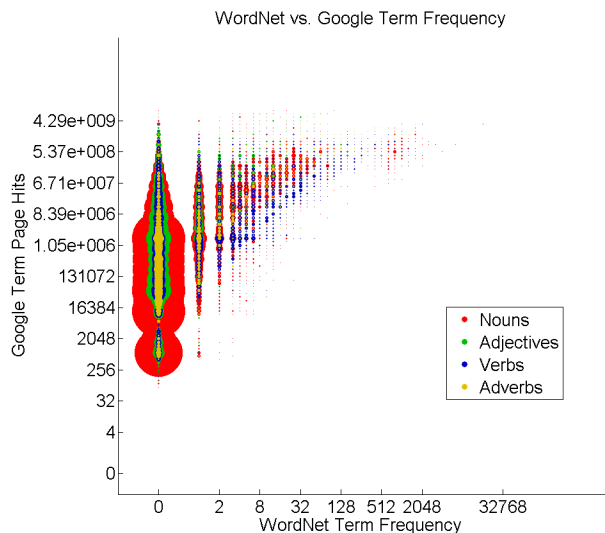


Figure 3: Comparison of term frequencies between WordNet/Brown and estimated Google page hits. The similarity in circle sizes apparent in bin zero is due to an upper limit Google imposes on its reported page hits.



Figure 4: Google search results include an estimate of document frequency over all indexed web pages.

part of speech and sense information, WordNet terms with multiple senses are assigned the same raw frequency score. Secondly, the frequency values are too large to sum up, in particular when approaching the root node for nouns. In order to avoid overflow, we need to scale all frequency counts without disturbing the resulting Information Content values. Specifically, we note that taking the square root of all frequency counts corresponds to halving all IC values, due to the logarithmic relationship between information content and frequency. We therefore scale frequency values by the square root before mapping them to WordNet senses.

Figure 2 compares the distribution of term frequencies in WordNet based on Brown to that derived from the large corpus of web documents. The large number of nouns and verbs distributed in bin zero of WordNet Term Frequency represent terms without any occurrence in the Brown corpus. Terms with frequency values in both corpora appear distributed in a wide band, which narrows as frequency increases.

3.2 Google Estimated Page Results

As an alternative to generating term frequencies based on large text corpora, we can use pre-computed page hits, such as those reported by the Google search engine (Figure 4). With constantly changing web documents, the estimated page frequency also changes frequently, although the differences are negligible. We

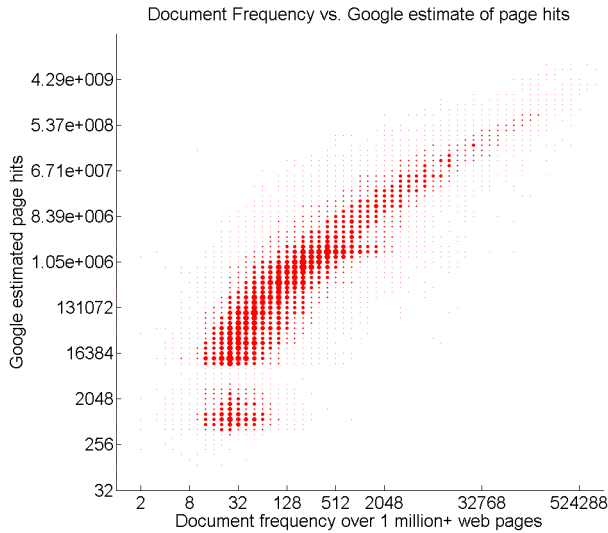


Figure 5: Document frequency in a set of 1,231,929 web pages versus Google page hit estimate. The two quantities are approximately linearly correlated, suggesting that one is a good predictor for the other.

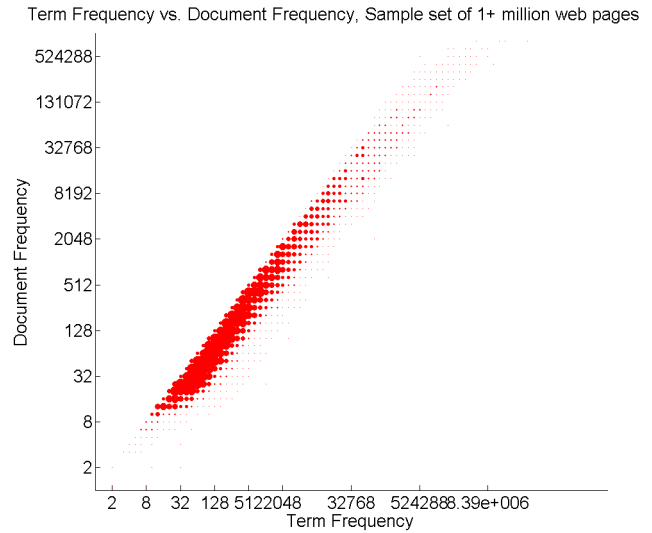


Figure 6: Term frequency versus document frequency for WordNet terms in a set of 1,231,929 web sites. The two quantities are linearly correlated, suggesting that one is a good predictor for the other.

Table 5: Top- and bottom-ranked document frequencies based on Google page hits.

| | | | |
|---------------------|--------------------|--------------------------------|---|
| www (8,730,000,000) | 15 (5,720,000,000) | giovanni vincenzo pecci (176) | left-handedness (143) |
| 1 (8,710,000,000) | 11 (5,600,000,000) | tendonous synovitis (176) | grey polypody (142) |
| 2 (8,200,000,000) | 12 (5,480,000,000) | st. baeda (175) | richard d. fosbury (131) |
| a (7,210,000,000) | 20 (5,430,000,000) | make vibrant sounds (169) | pubdental slit (131) |
| 3 (6,920,000,000) | 30 (5,180,000,000) | august f. mobius (166) | st. barbara's herb (112) |
| in (6,420,000,000) | 16 (5,150,000,000) | louis the bruiser (160) | micophage (111) |
| 5 (6,350,000,000) | 18 (4,880,000,000) | satyendra n. bose (160) | faggpting (111) |
| 10 (6,100,000,000) | 25 (4,860,000,000) | fluoxetine hydrochloride (157) | juan carlos victor maria de borbon y borbon (110) |
| 6 (6,030,000,000) | 7 (4,780,000,000) | saint ignatius' itch (156) | maxmilien de bethune (94) |
| 4 (5,880,000,000) | 21 (4,760,000,000) | forym (154) | fritz w. meissner (84) |

have generated a frequency table for all WordNet words and phrases using this readily available measure. Similar to the automated search and retrieval task presented in section 3.1, search queries are formulated with double-quoted WordNet entries to guarantee exact matches. The resulting frequency table (excerpt in Table 5) shows many similarities to the one described in section 3.1, although several web-centric terms, like “www”, “copyright”, and “html” appear in the top-ranked terms, which otherwise do not appear as frequent in standard text corpora. Figure 5 shows the similarity in term frequency distribution between estimated page hits computed by Google and page hits computed from a large sample of web pages. The near-linear relationship suggests that Google’s page hit estimate is a good approximation for most term document frequencies. The measure for information content for estimated page hits is computed from term frequencies as outlined in section 3.1. We find a similar comparison of frequency distributions between WordNet based on Brown and Google estimated page hits, with the exception that Google appears to impose an upper limit on its reported numbers (see Figure 3).

For both term frequency tables, we use page hits (i.e. document frequency) without taking into consideration term frequencies. Our analysis shows, however, that for WordNet terms, term frequency is linearly correlated to document frequency (see Figure 6). While this result may not hold true over the exhaustive set of all possible words and phrases, we focus exclusively on the WordNet database for which the results are significant.

4. CONCEPT-BASED VIDEO RETRIEVAL APPLICATION

We apply semantic similarity measures to video retrieval tasks by first mapping text search queries to WordNet and then mapping WordNet concepts to visual concepts. In the first step, our improved measure for information content increases the number of potential mappings between query terms and visual concepts, where previously some query terms could not be mapped due to the lack of IC values. The overall approach for concept-based query expansion and multimedia retrieval consists of building

Table 6: Excerpt from visual-to-semantic concept lexicon

| LSCOM Concept | Mapped WordNet concepts |
|--------------------|--|
| Address_Or_Speech | address#n#3, address#v#2 |
| Antenna | antenna#n#1 |
| Computer_TV-screen | computer#n#1, monitor#n#1, screen#n#3, tv#n#1, tv#n#2, laptop#n#1 |
| Court | court#n#1, court#n#2, court#n#4, judge#n#1, prosecution#n#2, defense#n#3, case#n#1, prosecute#v#1, prosecute#v#2, defend#v#6, rule#v#4 |

statistical detectors for a fixed set of semantic concepts, using the detectors to automatically tag the visual content with semantic labels, and then mapping arbitrary text queries to the pre-determined set of semantic tags. This method can be used to search visual content without any textual metadata, or it can be used to supplement alternative retrieval approaches, such as text-based, speech-based, or visual content-based retrieval. For more information on concept-based retrieval approaches, the reader is referred to [20-25]. For the experiments in this paper, we largely adopt the approach of [25] but instead of the Lesk similarity measure, we use the Jiang-Conrath similarity with Web-based IC values.

4.1 Automatic Concept Detection for Semantic Video Retrieval

As a first step we built support vector machine based semantic concept models [20] for all the annotated concepts of the LSCOM-lite lexicon based on visual features from the training collection. Each of these models can then be used to get a quantitative score indicating the presence of the corresponding concept in any test set video shot.

4.2 Mapping Query Terms to Visual Concepts

We have manually generated a lexicon, which maps LSCOM visual concepts to one or more WordNet concepts (see Table 6). A similar approach was used in [3], although the authors placed an upper limit on the number of WordNet concepts for any given visual concept. During a query analysis step, words and qualified WordNet phrases are extracted from the query. We then apply the Jiang-Conrath semantic similarity measure to all pairs of query terms and WordNet concepts in our lexicon to determine the closest-matching semantic concepts to the query. Selected visual concepts are then matched to visual detection results to create a ranked list of video shots.

4.3 Query Term Aggregation Approaches

Aggregating concept match scores across multiple query terms is an important aspect of the overall query-to-concept mapping approach. In general, the approach works by computing a similarity score between each query term and each visual concept. These scores are then aggregated across all query terms in order to compute an overall matching score between the query and each concept. The score aggregation method determines whether to

favor a concept that strongly matches a single query term or one that weakly matches multiple query terms. In general, this depends on the query itself, and whether the query terms are joined by OR logic or AND logic. AND logic is typically used to narrow the scope of the query, increase precision, and improve the relevancy of the top results, while OR logic is typically used to broaden the query, increase recall, and improve the coverage, or diversity, of the results. Unfortunately, users rarely indicate their intention but instead simply list their query terms. In such cases, most Internet search engines assume AND logic since recall is less of a concern with a large corpus such as the WWW. However, when searching in deep archives, or for exploratory-type queries, OR logic is a better choice. In this paper, we consider several aggregation strategies, which simulate various AND/OR logic flavors.

In particular, we use MAX similarity score aggregation to simulate Boolean OR combinations. In contrast, MIN score aggregation would correspond to Boolean AND combinations but we have found that strict AND logic is typically very harsh and provides very poor results for general queries. We therefore consider soft AND flavors, such as SUM and AVG of similarity scores. Both allow a matched concept’s weight to accumulate by soft-matching multiple query terms, where AVG also normalizes for query length (i.e., number of query terms), while SUM tends to favor longer queries. In the spirit of AND logic, both methods also “penalize” a concept that has a score of 0 with respect to any query terms since the overall SUM or AVG score will be lower. To lessen this penalty, we also consider a Non-Zero AVG aggregation method, which averages only the non-zero similarity scores, therefore allowing a concept to accumulate weight by matching multiple query terms while ignoring terms that have 0 matching scores. In terms of query scope, MAX allows for the broadest query interpretation, followed by Non-Zero AVG, followed by AVG, and finally SUM. As can be seen from the evaluation results, this order also roughly corresponds to the relative performance of the various aggregation methods on the TRECVID 2005-2007 corpora. This suggests that the TRECVID query topics are more recall-oriented as opposed to precision-oriented.

5. EVALUATION

We evaluate concept-based video retrieval with the new information content dataset on TRECVID 2005, 2006, and 2007 search tasks (see Figure 7). For TRECVID 2005 and 2006, we use a lexicon of 39 visual concepts matching the detected concepts (LSCOM-lite), while for TRECVID 2007, a larger set of 191 visual concepts was used. We also evaluate the relative performance of four query term aggregation approaches: MAX, SUM, AVG, and non-zero AVG. Results here are based on thresholding the final visual concept weights at mean + 1 standard deviation, computed over the concept vector [20].

We find that generally, a query term aggregation using a MAX approach outperforms the other tested measures. Comparing results based on information content from Google page hits and large sample set of web pages, we note various improvements over IC from Brown. For TRECVID 2005 data, IC from Google and sample web pages yield an improvement of 14% and 18%, respectively, and for TRECVID 2006, a significant improvement

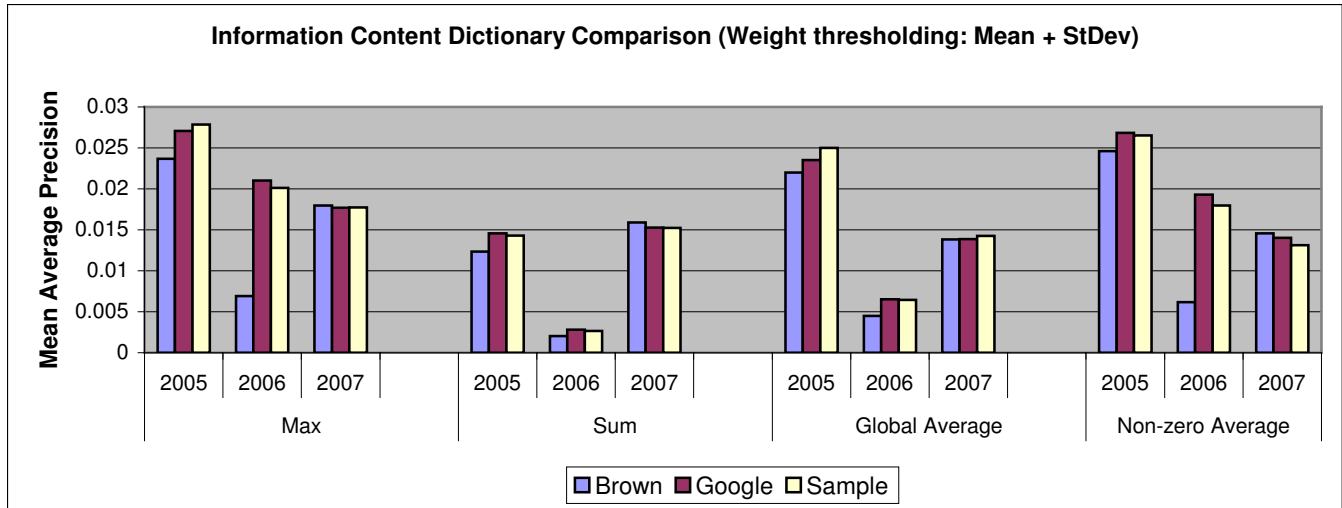


Figure 7: Mean Average precision results for TRECVID 2005, 2006, and 2007 search tasks, evaluated using four query term aggregation approaches over three information content sets (original Brown corpus, Google page hits, and large web page sample set). Generally, MAP scores improve when using more contemporary IC datasets. The aggregation approach of MAX also consistently outperforms the other methods.

of 204% and 191%. Results for TRECVID 2007 dropped only slightly by a statistically insignificant 2% and 1%. The large increases in mean average precision on TRECVID 2005-2006 data are due to the improved information content database, which more accurately accounts for the frequency of WordNet terms.

The recorded 1-2% drop in the TRECVID 2007 evaluation is the result of query 214 (“*very large crowd of people*”), which merits closer inspection. The three terms with semantic noun content (“*large*”, “*crowd*”, and “*people*”) have a Brown-based IC of (0, 4.66, and 2.22), respectively, and a web page sample set based IC of (3.99, 4.45, and 2.43). While the terms “*crowd*” and “*people*” have similar IC values across IC dictionaries, the term “*large*” is only captured in the new web-based IC. By including the term “*large*” in the WordNet-based Jiang-Conrath similarity measure, query 214’s average precision value drops, because the term “*large*” does not map as well to the visual concepts of “*crowd*” and “*people*”, and therefore introduces noise.

Query 195 (“*soccer goalposts*”) from the TRECVID 2006 dataset is a good example of a case, in which the new web-based IC provides a significant advantage over the Brown-based IC. Both of these terms are represented by information content of zero in the Brown-based IC due to a lack of occurrence in the original corpus. Using a semantic similarity measure such as Jiang-Conrath, Lin, or Resnik, neither term resolves in similarity to any other term in WordNet. With the new web-based database of information content, these terms were properly mapped to the visual concept of “*Sports*”, and search task precision increased from 0 to 0.34. While this is an extreme example, we find improvement in other queries, in which similarity to WordNet concepts may have been under-weighted due to rare usage at the time when the Brown corpus was introduced.

6. CONCLUSION

We have presented an approach for determining information content from two web-based sources, and demonstrated its

application to concept-based video retrieval. Information content is a frequently used measure, but its underlying database of term frequencies shows severe flaws due to outdated information from a too small text corpus. We show how two approaches of retrieving term frequency from web-based resources can be used to generate an improved information content database. We validate both approaches extensively, and find that the semantic similarity based on improved information content generally performs significantly better than the commonly used IC based on the Brown corpus.

7. ACKNOWLEDGEMENTS

This material is based upon work funded in part by the U.S. Government. Opinions, findings and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the U.S. Government.

8. REFERENCES

- [1] Fellbaum, C. WordNet: An Electronic Lexical Database. 1998. MIT Press, Cambridge, MA.
- [2] Zhai, Y., Liu, J., Shah, M. Automatic Query Expansion for News Video Retrieval. In Proceedings of the International Conference on Multimedia and Expo (Toronto, Canada, July 9-12, 2006). ICME '06. IEEE Press, New York, NY, 965-968.
- [3] Snoek, C.G.M., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., Worring, M. Adding Semantics to Detectors for Video Retrieval, IEEE Transactions on Multimedia, Vol. 9, Issue 5 (August 2007). IEEE Press, New York, NY, 975-986.
- [4] Wu, Z., Palmer, M. Verb semantics and lexical selection. In Proceedings of Annual Meeting of the Association for Computational Linguistics (Las Cruces, NM, June 27-30, 1994). Morgan Kaufmann, San Francisco, CA, 133-138.

- [5] Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceedings of the International Joint Conference on Artificial Intelligence (Montréal, Canada, August 20-25, 1995). IJCAI '95. Morgan Kaufmann, San Francisco, CA, 448-453.
- [6] Jiang, J.J., Conrath, D.W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In Proceedings of the International Conference Research on Computational Linguistics (Taipei, Taiwan, August 22-24, 1997). ROCLING X. 1997.
- [7] Lin, D. An information-theoretic definition of similarity. In Proceedings of the International Conference on Machine Learning (Madison, WI, 1998). ICML '98. Morgan Kaufmann, San Francisco, CA, 296-304.
- [8] Lesk, M.E. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the Special Interest Group Design of Communication Conference (Toronto, Canada, June 8-11). SIGDOC '86. ACM Press, New York, NY, 24-26.
- [9] Leacock, C., Chodorow, M., Miller, G.A. Using corpus statistics and WordNet relations for sense identification. In Computational Linguistics, Vol. 24, Number 1 (March 1998). MIT Press, Cambridge, MA, 147-165.
- [10] Over, P. Ianeva, T., Kraaij, W., Smeaton, A.F. TRECVID 2005 An Overview. In Proceedings of the NIST TRECVID 2005 Workshop (Gaithersburg, MD, November 14-15, 2005). TRECVID '05.
- [11] Over P., Ianeva, T., Kraaij, W., Smeaton, A.F. TRECVID 2006 Overview. In Proceedings of the NIST TRECVID 2006 Workshop (Gaithersburg, MD, November 13-14, 2006). TRECVID '06.
- [12] Over, P. Awad, G. Kraaij, W., Smeaton, A.F. TRECVID 2007 – An Introduction. In Proceedings of the NIST TRECVID 2007 Workshop (Gaithersburg, MD, November 5-6, 2007). TRECVID '07.
- [13] Pedersen, T., Patwardhan, Michelizzi, J. Wordnet::similarity – measuring the relatedness of concepts. In Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (Boston, MA, May 3-5, 2004). NAACL '04. Association for Computational Linguistics, Morristown, NJ, 38-41.
- [14] Patwardhan, S., Banerjee, S., Pedersen, T. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (Mexico City, Mexico, February 16-22, 2003). CILing '03. Springer Verlag, Berlin, Heidelberg, 241-257.
- [15] Seco, N., Veale, T., Hayes, J. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In Proceedings of the European Conference on Artificial Intelligence (Valencia, Spain, August 22-27, 2004). ECAI '04. IOS Press, Amsterdam, The Netherlands, 1089-1090.
- [16] Budanitsky, A., Hirst, G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Proceedings of the North American Chapter of the Association for Computational Linguistics Workshop (Pittsburgh, PA, June 2-7, 2001). NAACL '01. Association for Computational Linguistics, Morristown, NJ, 29-34.
- [17] Pucher, M. Performance Evaluation of WordNet-based Semantic Relatedness Measures for Word Prediction in Conversational Speech. In Proceedings of the International Workshop on Computational Semantics (Tilburg, Netherlands, January 12-14, 2005). IWCS 6.
- [18] Pedersen, T., Pakhomov, S. Developing Measures of Semantic Relatedness for the Biomedical Domain. Digital Technology Initiatives Forum (Minneapolis, MN, Feb 28, 2005). Digital Technology Center, University of Minnesota.
- [19] Naphade, M., Smith, J.R., Souvannavong, F. On the Detection of Semantic Concepts at TRECVID. In Proceedings of the ACM International Multimedia Conference (New York, NY, October 10-16, 2004). ACM Press, New York, NY, 660-667.
- [20] Natsev, A., Haubold, A., Tesic, J., Xie, L., Yan, R. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In Proceedings of the ACM International Conference on Multimedia (Augsburg, Germany, September 24-29, 2007). MM '07. ACM Press, New York, NY, 991-1000.
- [21] Neo, S.-Y., Zhao, J., Kan, M.-Y., Chua, T.-S.. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In Proceedings of the ACM International Conference on Image and Video Retrieval (Tempe, AZ, July 13-15, 2006). CIVR '06. Springer Verlag, Berlin, Heidelberg, 143-152.
- [22] Chang, S.-F., Hsu, W., Kennedy, L., Xie, L., Yanagawa, A., Zavesky, E., Zhang, D. Columbia University, TRECVID-2005 Video Search and High-Level Feature Extraction. In Proceedings of the NIST TRECVID 2005 Workshop (Gaithersburg, MD, November 14-15, 2005). TRECVID '05.
- [23] Chua, T.-S., Neo, S.-Y., Zheng, Y., Goh, H.-K., Xiao, Y., Zhao, M., Tang, S., Gao, S., Zhu, X., Chaisorn, L., Sun, Q. TRECVID-2006 by NUS-I2R. In Proceedings of the NIST TRECVID 2006 Workshop (Gaithersburg, MD, November 13-14, 2006). TRECVID '06.
- [24] Snoek, C. G. M., van Gemert, J. C., Geusebroek, J. M., Huurnink, B., Koelma, D. C., Nguyen, G. P., Rooij, O. D., Seinstra, F. J., Smeulders, A. W. M., Veenman, C. J., Worring, M. The MediaMill TRECVID 2005 Semantic Video Search Engine. In Proceedings of the NIST TRECVID 2005 Workshop (Gaithersburg, MD, November 14-15, 2005). TRECVID '05.
- [25] Haubold, A., Natsev, A., Naphade, M. Semantic multimedia retrieval using lexical query expansion and model-based reranking. In Proceedings of the International Conference on Multimedia and Expo (Toronto, Canada, July 9-12, 2006). ICME '06. IEEE Press, New York, NY, 1761-1764.
- [26] Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E.G.M., Milios, E.E. Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. In Proceedings of the ACM Workshop on Web Information and Data Management (Bremen, Germany, November 5, 2005). WIDM '05. ACM Press, New York, NY, 10-16.