SuMACC Project's Corpus

a Topic-based Query Extension Approach to Retrieve Multimedia Documents

Mohamed Morchid[†], Richard Dufour[†], Usman Niaz[‡], Francis Bouvier[◇], Clément de Groc[◇], Claude de Loupy[◇], Georges Linarès[†], Bernard Merialdo[‡], and Bertrand Peralta[∓*]

[†] LIA - University of Avignon, Avignon (France)
[◇] Syllabs, Paris (France)
[‡] EURECOM, Sophia Antipolis (France)
[∓] WIKIO, Paris (France)
{mohamed.morchid,richard.dufour,georges.linares}@univ-avignon.fr
{bouvier,groc,loupy}@syllabs.com
{usman.niaz,bernard.merialdo}@eurecom.fr
{bertrand.peralta}@ebuzzing.com

Abstract. The SuMACC project aims at automatically tracking new multimodal entities on Internet. The goal of the project is to propose robust multimedia methods that define relevant patterns allowing to automatically retrieve these entities. This paper describes the SuMACC corpus collected on video-sharing platforms using word-queries. Since concepts are limited to a single or few words, querying video-sharing platforms with the concept only can easily introduce irrelevant collected videos. In this paper, we propose to use an extended query obtained by mapping the initial concept into a topic space from a Latent Dirichlet Allocation (LDA) algorithm. This topic-based query extension approach allows to better retrieve videos related to the targeted concept. As a result, a corpus of 7,517 videos, extracted using the simple (*i.e.* concept only) and the extended queries, from 47 concepts, was obtained. Results show the effectiveness of the proposed thematic querying approach compared to the simple concept query in terms of relevance ($+21\%$) and ambiguity (-4%). The annotation process as well as the corpus statistics are detailed in this paper.

Keywords: Multimedia corpus, Annotation, Latent Dirichlet Allocation, Topic modeling, Extended queries

1 Introduction: the SuMACC Project

The search of a concept in multimedia database or on the Internet encounters major issues due to the diversity of concept representations, that may depend on one or several different modalities, such as pictures, video, speech, text, sounds…Typically, a concept such as *olympic games* may be mapped into videos of opening ceremony, in

* This work was funded by the SuMACC project supported by the French National Research Agency (ANR) under contract ANR-10-CORD-007.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 29–36, 2014.

text documents focusing on a specific race, in radio shows…The SuMACC aproject[1] aims to develop models supporting these variabilities related to the multimedia contents, with a particular focus on the Web.

Methods for concept discovery and tracking in text documents have been largely studied in the last decades. These methods are now relatively mature and effective. Moreover, the video processing community produced great efforts to design methods for tracking concrete objects or object categories, especially in the TrecVid evaluation campaigns [8].

Nonetheless, multimodal approaches remain poorly developed, most of previous works proposing solutions for only one modality (video, audio or text). From a technological point of view, most of the identification methods are based on statistical models. To correctly estimate model parameters, a large amount of data is however mandatory. Collecting and annotating such large corpus is generally too costly, thus avoiding the emergence of multimedia approaches.

The SuMACC project addresses these two major issues related to the multimodal representation of concepts and to the training strategies that could enable a low-cost estimate of concept signatures. This paper describes the collected and annotated corpus used to propose multimodal searches and training strategies.

The next section presents the method followed to collect data, especially the query strategy based on a topic-representation of concepts. Section 3 describes the obtained corpus and its annotation protocol as well as a discuss about size, nature and quality of the collected database. Section 4 concludes and presents some future works.

2 Collecting Evaluation Data: Methodology

2.1 Motivation and Principle

To evaluate concept retrieval methods, plausible scenarios have to be simulated. These simulations require a set of realistic queries (*i.e.* that could be asked by users), a large video database in which targeted concepts will be searched, and tags that indicate if videos effectively contain the targeted concepts.

One of the major difficulty in collecting such a corpus is due to the fact that the data set is basically composed of videos obtained by requesting the search engines (SE) of video sharing platforms. The implicit video tagging performed by the SE can not be used as a ground truth: firstly, the SE makes errors; secondly, the collected database should be designed to enable simulation of a realistic information retrieval tasks. Consequently, the collected set has to contain not only the videos having the targeted concepts, but also ambiguous ones. Our proposal is to use a query extension method that allows the hit of videos related to the targeted concept and ambiguous ones.

The query generation process starts from the initial concept characterization. Each concept is expressed as a keyword (or a key expression) that could be used as a query to obtain concept-related videos. We expand this primary query by using closed keywords. In the context of information retrieval, most of the previous works proposed to expand the initial user query by using a vector space model [4], a word similarity matrix [6], the

[1] `http://sumacc.univ-avignon.fr/`

Table 1. Statistics on the French Wikipedia dump for the topic space training.

Characteristic	Statistic
Number of articles	3, 197, 395
Number of words ($\|D\|$)	898, 645, 071
Number of unique words (N)	13, 182, 180
Number of words per article	281.05
Number of unique words per article	4.12

analysis of social networks [2] or visual descriptors [5]. We propose a query extension by mapping the concept into a topic space obtained by a Latent Dirichlet Allocation (LDA) method, presented in details in next section. By using such a topic space for query extension, we aim to introduce in the dataset not only negative examples, but ambiguous examples corresponding to ambiguous semantic contents.

2.2 LDA Approach

Latent Dirichlet Allocation (LDA) [3] is a generative probabilistic model which considers a document as a *bag-of-words* produced by a set of latent topics. Word occurrences are linked by latent variables that determines the distribution of topics in a document. This decomposition model of documents offers good generalization abilities compared to other generative models that are commonly used in automatic language processing such as Latent Semantic Indexing (LSI) or Probabilistic Latent Semantic Indexing (PLSI) [1,7].

A topic z, associated with a LDA class, is represented by a vector V_z, whose coefficients represent the probability of words w_i knowing the topic z:

$$V_i^z = P(w_i|z)$$

This method requires a large dataset to build a global model. Our training corpus D is composed by documents from the French Wikipedia dump (see Table 1), containing about 900 million words.

2.3 Building Query Set

Videos are collected by querying the Syllabs multimedia fetcher, that allows to search videos from four video sharing platforms: DailyMotion, Youtube, Vimeo and Flickr. The queries are composed of the initial concept (*i.e.* one keyword) and a set of extended keywords obtained with a LDA-based technique; this method consists in identifying the *n-best* words of the closest topic of the concept. The set of n words is considered as the first expanded query. This first extension step is followed by a second one, where we use the first expanded query to select the *k-best* concepts. Finally, a set of m words are extracted by cumulating conditional word probabilities belonging to the m best LDA classes. This step yields to get the m most relevant words for each initial concept. This second expanded query is submitted to the Syllabs multimedia fetcher to collect videos matching the final query. Therefore, the document retrieval process performs 4 steps:

– building an off-line thematic representation;
– mapping the concept into the topic space to extract the best topic z. Then, m words of the topic z are chosen to compose a first expanded query;
– mapping this first expanded query into the topic space to find its k closest topics. Then, scoring each word of $\mathbf{V}$ to find the final *expanded query*;
– sending each expanded query (and the initial one) to a multimedia fetcher to retrieve

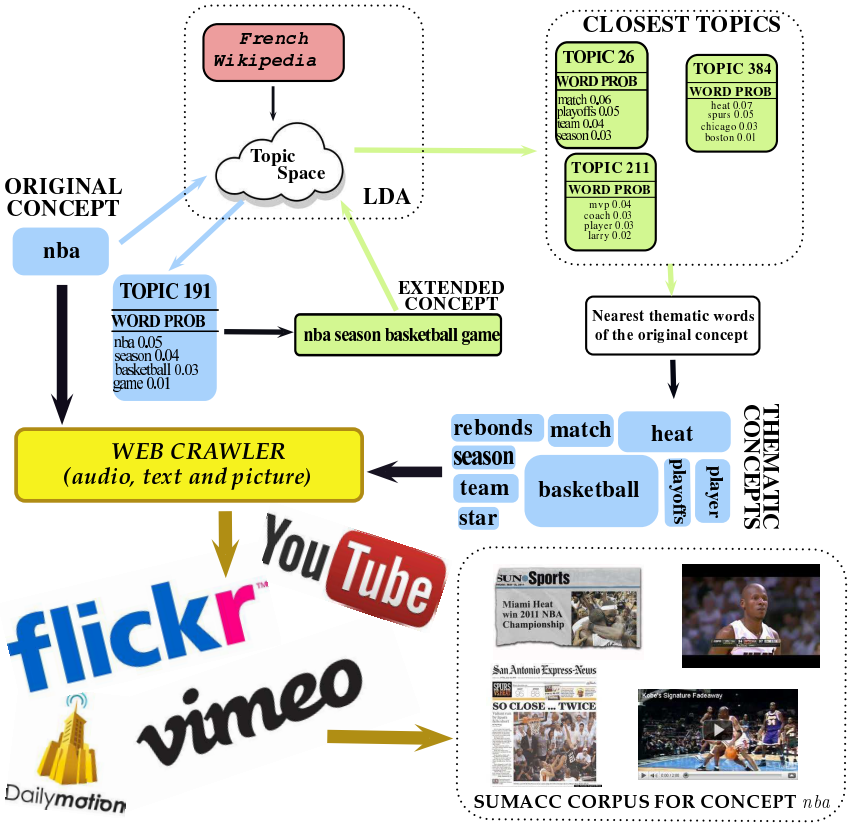

Fig. 1. Example of the initial query *nba* and its extension to retrieve documents.

Figure 1 shows an example of an initial concept *NBA* processed in the document retrieval system (audio, text or picture documents).

3 The SuMACC Corpus

The manual corpus annotation consists in checking, in each video, the presence of the concept which was used to collect it (via the expanded queries).

The SuMACC concept list is composed of 47 concepts corresponding to different kinds of entities that may be searched on the Web, with respect to the project goals;

therefore, some of them are very concrete (such as *Ipad2* or *Jennifer Lopez*), while others are much more abstract (such as *racism*). In Table 2, the number of videos retrieved by the initial queries and their expanded version is compared. We observe that the initial query allowed to retrieve 1,574 (21%) videos, the 5,943 (79%) others being extracted thanks to the thematic queries. Some concepts such as *can_2012* or *ipad_2* allow to retrieve more documents. This is mainly due to their popularity and to the fact that these precise concepts can better describe a video than a more general one such as *renseignement et espionnage (intelligence and espionage)* or *âge de départ à la retraite (retirement age)* which can be associated to a lot of heterogeneous videos.

The time duration of all documents related to a concept is detailed in Table 2. This table shows that the time duration is well distributed among the 47 concepts. The total number of hours of the corpus is about 89 days. This represents 2,162 hours of videos. Each concept contains 329 videos in average.

With this corpus of video documents, a set of video-related text documents that describes the extracted videos is added to the SuMACC corpus. Note that a description is not available for all the videos. Thus, 1,410 descriptions are collected for the 7,517 videos of the corpus. This is a real context fact: few videos have a textual document to describe their content. This set of text documents contains 9,692 unique words for a total of 56,474 running words.

As expected, Table 3 shows that the videos retrieved with the thematic queries are globally considered more relevant that the ones from the initial queries. Nonetheless, the videos retrieved with the thematic queries have a tendency to be much more variable than the one obtained with the initial queries. Indeed, if we take a look at the column *Ambiguous* of Table 3, we can notice that the total proportion of ambiguous videos is about 14% for the original queries, while only 10% is for the thematic queries.

Moreover, the proportion of videos considered as relevant (bold in column *Yes* in Table 3) is higher with extended queries than the initial ones (query containing the concept only). In details, 9 concepts among the 47 ones, have a proportion of relevant videos beyond 80% with the use of a topic space representation (only 2 for the not-extended queries).

The main issue of this video collecting task is to obtain a sufficient variability in the video content while remaining close to the query topic.

4 Conclusion

In this paper, the SuMACC project corpus was described as well as an unsupervised method to retrieve a large set of concept-related multimedia documents. This method expands simple requests in order to get a realistic sampling of videos returned by a search engine. Query extension is based on a 2-step mapping of keywords into a topic space estimated by a LDA approach. This method allows to add a necessary variability in the corpus while respecting a realistic ambiguity due to topic proximity. As a result, up to 23,000 videos from 47 concepts is obtained, 1,432 of them being annotated in a first annotation campaign. The corpus will be freely available under GPL license by the end of 2014.

Table 2. Statistics on the concepts of the SuMACC project.

Concept	#documents	%	time	%
accident_de_la_route (*road_accident*)	535.0	7.117	1D-5:12:59	5.827
age_de_départ_à_la_retraite (*retirement_age*)	30.0	0.399	0:40:33	0.135
alpes	33.0	0.439	2:31:21	0.503
apple	114.0	1.517	7:0:45	1.399
applications_iphone	206.0	2.74	9:53:37	1.973
barack_obama	455.0	6.053	1D-6:49:42	6.148
barcelone-real_madrid	32.0	0.426	4:2:38	0.807
bnp_paribas	30.0	0.399	0:56:5	0.186
brad_pitt	455.0	6.053	1D-0:58:3	4.98
can_2012	30.0	0.399	1:50:56	0.369
cisjordanie	30.0	0.399	0:19:23	0.064
consoles_portables (*portable_game_consoles*)	30.0	0.399	0:56:24	0.187
cosmétique (*cosmetic*)	843.0	11.215	3D-2:38:17	14.886
dominique_strauss-kahn	341.0	4.536	14:29:41	2.891
françois_hollande	144.0	1.916	7:59:25	1.594
fukushima	33.0	0.439	1:14:43	0.248
galeries_d'art (*art_galleries*)	686.0	9.126	1D-10:52:20	6.955
gameplay	30.0	0.399	10:33:54	2.107
ground_zero	154.0	2.049	10:57:52	2.187
hôtel_de_ville (*city_hall*)	30.0	0.399	0:40:34	0.135
ipad_2	76.0	1.011	5:42:23	1.138
jacques_chirac	349.0	4.643	20:19:10	4.053
javier_pastore	30.0	0.399	0:58:44	0.195
jennifer_lopez	383.0	5.095	1D-4:49:39	5.749
kanye_west	38.0	0.506	0:33:22	0.111
liberté_d'expression (*freedom_of_expression*)	30.0	0.399	1:37:19	0.323
londres_2012	39.0	0.519	1:44:50	0.348
marché_financier (*financial_market*)	38.0	0.506	1:46:10	0.353
mouammar_kadhafi	112.0	1.49	6:6:38	1.219
nba	51.0	0.678	6:0:19	1.198
oscars	36.0	0.479	1:12:25	0.241
otan	30.0	0.399	2:58:42	0.594
paris_saint-germain	33.0	0.439	0:51:59	0.173
prix_nobel_de_la_paix (*nobel_peace_prize*)	42.0	0.559	1:19:41	0.265
présidentielle_2012 (*presidential_elections_2012*)	80.0	1.064	2:27:14	0.489
psn	30.0	0.399	2:47:37	0.557
racisme (*racism*)	155.0	2.062	11:44:10	2.341
real_madrid	30.0	0.399	0:24:7	0.08
renseignement_et_espionnage (*intelligence_and_espionage*)	34.0	0.452	3:50:59	0.768
semaine_de_la_mode (*fashion_week*)	568.0	7.556	2D-1:46:16	9.926
stade_de_france	30.0	0.399	4:38:35	0.926
steve_jobs	343.0	4.563	1D-0:54:10	4.967
tournages (*filming*)	30.0	0.399	0:41:0	0.136
vernissages_et_expositions (*openings_and_exhibitions*)	352.0	4.683	22:51:59	4.561
vitrolles	30.0	0.399	1:22:13	0.273
washington	180.0	2.395	20:49:14	4.153
zone_euro (*eurozone*)	127.0	1.69	6:25:40	1.282
Total	**7,517**	–	**20D-21:23:47**	–

Table 3. Comparison between the initial and the expanded queries after manual annotation.

	Initial Queries			**Expanded Queries**		
Concept	Yes	No	Ambiguous	Yes	No	Ambiguous
accident_de_la_route (*road_accident*)	50	44	4	49	41	9
age_de_départ_à_la_retraite (*retirement_age*)	66	16	16	42	19	38
alpes	21	71	7	55	44	0
apple	6	86	6	17	68	13
applications_iphone	30	65	3	62	34	3
barack_obama	73	24	2	63	17	18
barcelone-real_madrid	11	77	11	30	25	45
bnp_paribas	16	83	0	68	31	0
brad_pitt	69	18	11	47	39	12
can_2012	33	50	16	12	87	0
cisjordanie	60	20	20	96	4	0
consoles_portables (*portable_game_consoles*)	20	60	20	**90**	9	0
cosmétique (*cosmetic*)	**87**	11	0	60	20	18
dominique_strauss-kahn	27	63	9	**88**	4	6
françois_hollande	66	9	23	74	17	8
fukushima	11	88	0	66	33	0
galeries_d'art (*art_galleries*)	40	31	28	54	28	16
gameplay	66	0	33	**96**	0	3
ground_zero	13	82	4	36	56	7
hôtel_de_ville (*city_hall*)	20	40	40	8	84	8
ipad_2	50	40	10	25	53	21
jacques_chirac	63	32	4	76	10	13
javier_pastore	20	70	10	15	78	5
jennifer_lopez	42	53	4	57	30	12
kanye_west	33	33	33	**82**	7	10
liberté_d'expression (*freedom_of_expression*)	33	44	22	64	29	5
londres_2012	30	46	23	64	20	16
marché_financier (*financial_market*)	25	66	8	70	20	8
mouammar_kadhafi	31	52	15	**81**	15	3
nba	33	0	66	65	18	16
oscars	25	62	12	44	55	0
otan	40	20	40	68	28	4
paris_saint-germain	16	50	33	**88**	7	3
prix_nobel_de_la_paix (*nobel_peace_prize*)	28	57	14	75	25	0
présidentielle_2012 (*presidential_elections_2012*)	14	57	28	54	33	12
psn	33	50	16	**91**	4	4
racisme (*racism*)	**89**	3	7	61	27	11
real_madrid	13	80	6	20	66	13
renseignement_et_espionnage (*intelligence_and_espionage*)	71	14	14	**80**	19	0
semaine_de_la_mode (*fashion_week*)	0	94	4	44	34	20
stade_de_france	25	62	12	61	28	9
steve_jobs	10	87	2	51	36	11
tournages (*filming*)	16	66	16	**83**	12	4
vernissages_et_expositions (*openings_and_exhibitions*)	64	10	25	76	10	12
vitrolles	28	42	28	21	65	13
washington	10	86	3	27	42	29
zone_euro (*eurozone*)	58	27	13	46	43	9
Total	**37%**	**47%**	**14%**	**58%**	**30%**	**10%**

References

1. Ando, R.K., Lee, L.: Iterative Residual Rescaling: An analysis and generalization of LSI. In: Proceedings of SIGIR. pp. 154–162 (2001)
2. Bertier, M., Guerraoui, R., Leroy, V., Kermarrec, A.M.: Toward personalized query expansion. In: ACM EuroSys Workshop on Social Network Systems (SNS). pp. 7–12 (2009)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
4. Crouch, C.J., Crouch, D.B., Nareddy, K.R.: The automatic generation of extended queries. In: ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 369–383 (1990)
5. Feng, B., Cao, J., Chen, Z., Zhang, Y., Lin, S.: Multi-modal query expansion for web video search. In: ACM SIGIR conference on Research and development in Information Retrieval. pp. 721–722 (2010)
6. Gauch, S., Wang, J., Rachakonda, S.M.: A corpus analysis approach for automatic query expansion and its extension to multiple databases. ACM Transactions on Information Systems 17(3), 250–269 (1999)
7. Hofmann, T.: Probabilistic latent semantic indexing. In: ACM SIGIR conference on Research and development in Information Retrieval. pp. 50–57 (1999)
8. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM international workshop on Multimedia information retrieval. pp. 321–330 (2006)