RGU at ImageCLEF2010 Wikipedia Retrieval Task

Jun Wang, Dawei Song, and Leszek Kaliciak

School of Computing, The Robert Gordon University, Aberdeen, UK
{j.wang3,d.song,l.kaliciak}@rgu.ac.uk

Abstract. This working notes paper describes our first participation in the ImageCLEF2010 Wikipedia Retrieval Task[1]. In this task, we mainly test our Quantum Theory inspired retrieval function on cross media retrieval. Instead of heuristically combining the ranking scores independently from different media types, we develop a tensor product based model to represent textual and visual content features of an image as a non-separable composite system. Such system incorporates the statistical/semantic dependencies between certain features. Then the ranking scores of the images are computed in a way as quantum measurement does. Meanwhile, we also test a new local feature that we have developed for content based image retrieval.

1 Introduction

In our participation to ImageCLEF we have submitted runs in the Wikipedia Retrieval Task, which are based on our Quantum Theory inspired retrieval model. This model applies tensor product to represent the textual and visual features of an image as a n-order tensor in order to capture the non-separability of textual and visual features. The order of the tensor depends on the visual features that are going to be incorporated in the image retrieval.

We have tested this retrieval model on the ImageCLEF2007 data collection that has 20,000 images in total, and achieved some promising results. Therefore we would like to test it on a larger scale dataset.

2 Tensor Based Retrieval Model

Mathematically, the tensor product is used to construct a new vector space or a new tensor, where the relationship of the vector spaces can be expressed. In quantum mechanics, the tensor product can be used to expand Hilbert spaces or construct a composite system with separate systems.

Suppose an image represented in a single feature space to be a single system S_i, then that image can be constructed as a composite system by tensor product the separated systems from different feature spaces:

$$S = S_1 \otimes S_2 \cdots \otimes S_n \tag{1}$$

Where S_i is a system in a single feature space.

Next lets look at how we present a single system in the Hilbert space with Dirac notation. Here we will not introduce the detail of Dirac notation, readers who are interested can refer to Van Rijsbergen's book "The Geometry of Information Retrieval"[2].

With quantum mechanics, traditionally vector based representation of documents are represented as superposed states. The textual feature of an image is:

$$|T\rangle = \sum_i w_{t_i} |t_i\rangle \tag{2}$$

where $\sum_i w_{t_i}^2 = 1$, and the amplitude w_{t_i} is proportional to the probability that the document is about the term t_i. w_{t_i} can be set up with any term weighting scheme, and we adopted TF-IDF in our experiment.

Similarly, the histogram of content feature can also be represented as a superposed state in a content feature space:

$$|F\rangle = \sum_i w_{f_i} |f_i\rangle \tag{3}$$

where $\sum_i w_{f_i}^2 = 1$, f_i is the a particular feature bin, and w_{f_i} is proportional to the number of pixels falling into the corresponding bin of the feature space.

When more than one features are used to represent an image, the representation will be:

$$|D\rangle = |T\rangle \otimes |F_1\rangle \otimes |F_2\rangle \cdots \otimes |F_n\rangle \tag{4}$$

In our pilot study, we only combine one content feature with the textual feature.

$$\begin{aligned} |D\rangle &= |T\rangle \otimes |F\rangle && (5)\\ &= \sum_{ij} \gamma_{ij} |t_i \otimes f_j\rangle && (6) \end{aligned}$$

Here, i and j are the dimensionalities of textual and content features. When textual and content feature are completely independent, then $\gamma_{ij} = w_{t_i} \cdot w_{f_j}$. However, this does not hold generally. Extra operation is necessary to reflect the non-separability of the two features. It can be operationalised as the co-occurrence or correlation of the features. The tensor product enables the expansion of feature spaces in a seamless way and incorporates the correlations between the feature spaces.

With superposed representation, the similarity of a document and a query can be viewed as the probability that the document projects onto the sub-space that is expanded by the query features.

The probability that a document collapses to a state is:

$$P(t_i|d) = |\langle t_i|T\rangle|^2 = w_{t_i}^2 \tag{7}$$

It can be described as a projection onto a space spanned by $|t_i\rangle$:

$$P(t_i|d) = \langle t_i|\rho_d|t_i\rangle = w_{t_i}^2 \tag{8}$$

where $\rho_d = |d\rangle\langle d|$ is the density matrix of document d, and $w_{t_i}^2$ is the probability that the term t_i appears in the document or the probability that the document is about the term.

With the textual and visual composite system, the density matrix of a document is:

$$\rho_D = |D\rangle\langle D| = \sum_{ij} \gamma_{ij}^2 |t_i f_j\rangle\langle t_i f_j| \tag{9}$$

Then the similarity between a document and a query is:

$$sim(d, q) = tr(\rho_d \rho_q) \tag{10}$$

For a visual feature, each of its dimensions can be treated as orthogonal to other dimensions. Because the blue color pixel can never be counted as a red pixel. While for textual feature, two terms can be semantically related, e.g. cup and mug may refer to the same thing in one document. To represent the document with orthogonal textual basis, we can use a transformation matrix to fulfill the requirement:

$$\begin{align} \rho_d &= \sum_i w_i^2 |t_i\rangle\langle t_i| \tag{11} \\ &= \sum_i w_i^2 \sum_j U_{ij} |e_j\rangle \sum_k \langle e_k| U_{ik} \tag{12} \\ &= \sum_{jk} \rho_{jk} |e_j\rangle\langle e_k| \tag{13} \end{align}$$

In the current experiment, we assume that all the terms are orthogonal to simplify the calculation, then:

$$\begin{align} sim(d, q) &= tr(\sum_i d_i^2 |t_i\rangle\langle t_i| Q) \tag{14} \\ &= tr(\sum_i d_i^2 |t_i\rangle\langle t_i| \sum_j q_j^2 |t_j\rangle\langle t_j|) \tag{15} \\ &= \sum_i d_i^2 q_i^2 \tag{16} \end{align}$$

3 Experiment Settings and Results

3.1 Text Processing

When we associated texts with images, we not only used annotation documents, but also used Wikipedia pages, hoping the original full text document can provide more semantic information.

Although in the Wikipedia task, the images may be annotated by more than one language, due to our language expertise we only parse the English language. If the images are not annotated by English, or do not appeared in English Wikipedia page, then this image will not be indexed. It also means that this image will never be retrieved in the runs using both text and content features as queries.

For the annotation files, we parsed all the terms from name, description, comment and caption entries if they contain any term. For the Wikipedia dump files, We parsed all the terms in the Wikipedia page, including the the link names within the page. However we do not go further into the linked page. The tags particular to the Wikipedia webpage are removed.

Same for the queries, we only use their English titles during the retrieval stage.

3.2 Content Feature

Apart from the content features provided by the organizer, we also used our local feature, which is based on the "bag of features"/"bag of visual words" approach. The feature extraction consists of following stages: image sampling, description of local patches, feature vector construction, visual dictionary generation and histogram computation. The number of sample points is 900 per image and the sampling is purely random. We open a square window - local patch (10 by 10 pixels wide) centred at the sample point. Each local patch is represented in a form of three colour moments computed for individual colour channels in HSV colour space. Thus obtained vector representation of a local patch has 9 dimensions. We apply the K-means clustering to the training set in order to obtain the codebook of visual words. Finally, we create a histogram of visual word counts by calculating manhattan distance between image patches and cluster centroids and generate a vector representation for each image from the collection. Thus obtained vector representation of an image has 40 dimensions.

3.3 Experiment runs and Results

With the textual and visual feature available, we submitted the following runs, some of which are based on content feature only and some are content and textual feature mixed.

- Text and content mixed retrieval
  - T+F_L : retrieve on annotation first, then re-rank with our local feature
  - T+F_C : retrieve on annotation first, then re-rank with cime
  - TXF_L : quantum-like measurement on tensor product space of annotation vector and our local feature vector
  - TXF_C : quantum-like measurement on tensor product space of annotation vector and feature cime vector
  - combine: T+F_C based retrieval. When the length of result from T+F_C is less than 1000, then the images from content retrieval are appended into the result list.
  - W+F_C : retrieve on Wikipedia file first, then re-rank with cime
- Content only retrieval
  - c_leszek: city block distance with our new local feature
  - c_add: city block distance with all content feature provided by ImageCLEF organiser

For the mixed retrieval, we did not run the mixed retrieval process on the whole collection due to the huge size. We ran the text retrieval first, then applied mixed retrieval to the re-ranking.

From table 1, we can see that the retrieval result based on content feature has extremely low MAP. The retrieval results from text and content mixed retrieval in ImageCLEF2010 is also considerably lower than our results from ImageCLEF2007 whose MAPs that are around 14%.

After further looking into the tensor product based experiments, we did find a bug in the code. Because a "+" operator is missing, which resulted in that only the last textual and content feature dimension had been used to re-rank the image. This accounts the poor performance of the submission runs.

We have corrected the code and re-run the experiments. The text based result is MAP=0.0939 and P@10=0.3485; The tensor product based result with our local feature is MAP=0.0665 and P@10=0.2000. There is a slightly improvement after removing the bug from the experiment, but it is still a lot of worse than text based retrieval, which needs a further investigation.

Run	Modality	field(s)	MAP	P@10
combine	Mixed	TITLEIMG	0.0617	0.2271
T+F_L	Mixed	TITLEIMG	0.0617	0.2257
TXF_C	Mixed	TITLEIMG	0.0486	0.1443
TXF_L	Mixed	TITLEIMG	0.0484	0.1500
T+F_C	Mixed	TITLEIMG	0.0325	0.1143
W+F_C	Mixed	TITLEIMG	0.0031	0.0086
c_leszek	Visual	IMG	0.0069	0.0614
c_add	Visual	IMG	0.0003	0.0100

Table 1. Our runs in ImageCLEF2010

4 Conclusions and Future Works

In this notes paper, we reported our quantum theory inspired multimedia retrieval framework, which provides a formal and flexible way to expand the feature spaces, and seamlessly integrate different features. The similarity measurement between query and document follows quantum measurement.

We did not include the correlations between dimensions across different feature spaces in this year submission. We would like to investigate such issue in the continuing study, and further the study with entanglement concept in quantum mechanics.

In the current experiments, we assumed each word is orthogonal, which does not hold and can be relaxed in the future. We can solve this problem with either dimensionality reduction or utilize the thesaurus to remove the synonyms. This will also facilitate the ranking computation.

Acknowledgements

This work is funded in part by the UK's Engineering and Physical Sciences Research Council (grant number: EP/F014708/2).

References

1. Adrian Popescu, Theodora Tsikrika, and Jana Kludas. Overview of the wikipedia retrieval task at imageclef 2010. In *Working Notes of CLEF 2010*, Padova, Italy, 2010.
2. C. J. van Rijsbergen, editor. *The Geometry of Information Retrieval*. Cambridge University Press, 2004.