Bridging the Semantic Gaps in Information Retrieval : Advanced Image Search Using Topic Model.

著者	Nguyen Cam Tu
号	17
学位授与機関	Tohoku University
学位授与番号	情博第518号
URL	http://hdl.handle.net/10097/59886

氏名（本籍地）	ウェン　キャン チュウ Nguyen Cam Tu
学位の種類	博　士（情報科学）
学位記番号	情　博　第518号
学位授与年月日	平成23年9月15日
学位授与の要件	学位規則第4条第1項該当
研究科、専攻	東北大学大学院情報科学研究科（博士課程）システム情報科学専攻
学位論文題目	Bridging the Semantic Gaps in Information Retrieval : Advanced Image Search Using Topic Model.（情報検索における意味的ギャップの解消：トピックモデルを用いた先進的画像探索）
論文審査委員	（主査）東北大学教授　徳山　豪 東北大学教授　篠原　歩　東北大学教授　乾　健太郎 東北大学准教授　全　眞　嬉

論文内容の要旨

Chapter 1: Introduction.

The overall goal of this thesis is to bridge semantic gaps in information retrieval especially image retrieval. We consider two types of semantic gaps, that are the gap between textual captions and human concepts, and the gap between image features and textual captions. The first gap is related to multiple linguistic phenomenon such as synonymy, polysemy. It causes data mismatching, reduces retrieval performance in text-based searching. The latter is between visual features of images and descriptive labels and prevents object recognition extending to a larger number of objects. Toward advanced context-based image search, we focus on closing the gaps using topic models.

Chapter 2: Hidden Semantic and Topic Analysis

This chapter presents methodologies in semantic representation such as semantic networks, semantic space, hidden topics as well as the relationships among them. Over the years, semantic representation is an active topic in artificial intelligence, machine learning, data mining, etc. as well as a matter for debate in cognitive psychology. In order to make machines "more intelligent" and bridge the "semantic gap", computer scientists are interested in studying how humans perceive semantic concepts. Three typical approaches to semantic representation are Semantic Network, Semantic Space, and Topic Models.

Semantic network presents concepts by nodes and relationships between concepts are encoded by edges. Semantic network usually be hand-coded by analyzing the domain of interest and represented by ontology. Wordnet is one famous example of this type.

In semantic space, words are represented as points in Euclidean space and proximity implies semantic association. This is the solution produced by Latent Semantic Analysis (LSA), which is sometimes referred to as Latent Semantic Indexing (LSI) in the context of its application in Information Retrieval.

This approach of topic models is based on the idea that documents are mixture of topics and each topic is a probability distribution over words. Although topic models also aim at semantic representation and dimensionality reduction as LSA, their approach is in the view of statistically generative models instead of vector space. Probabilistic Latent Semantic Analysis (pLSA) is the pioneer in this approach. Latent Dirichlet Allocation (LDA) was successively proposed as a more complete generative model compared to pLSA, this topic model has received more and more attentions with applications in multiple fields including text and image retrieval.

Chapter 3: Web Search Clustering and Labeling with Hidden Topics

Although the performance of search engines is enhanced day by day, it is a tedious and time-consuming task to navigate through hundreds to hundred thousands of "snippets" returned from search engines. A study of search engine logs argued that *"over half of users did not access result beyond the first page and more than three in four users did not go beyond viewing two pages"*. Since most of search engines display from about 10 to 20 results per page, a large number of users is unwilling to browse more than 30 results. One solution to manage that large result set is clustering. Like document clustering, search results clustering groups similar "search snippets" together based on their similarity; thus snippets relating to a certain topic will hopefully be placed in a single cluster. This can help users locate their information of interest and capture an overview of the retrieved results easily and quickly. In contrast to document clustering, search results clustering needs to be performed for each query request and be limited to the number of results returned from search engines. In contrast to normal documents, these snippets are usually noisier, less topic-focused, and much shorter, that is, they contain from a dozen words to a few sentences. Consequently, they do not provide enough shared-context for good similarity measure.

This chapter introduces a general framework for clustering and labeling with hidden topics discovered from a large-scale data collection. This framework is able to deal with the shortness of snippets as well as provide better topic-oriented clustering results. The underlying idea is that we collect a large collection, which we call the "universal dataset", and then do topic estimation for it based on recent successful topic models such as pLSA, LDA. It is worth reminding that the topic estimation needs to be done for a large corpus of long documents (the universal dataset) so that the topic model can be more precise. Once the topic model has been converged, it can be considered as one type of linguistic knowledge which captures the relationships between words. Based on the converged topic model, we are able to perform topic inference for (short) search results to obtain the intended topics. The topics are then combined with the original snippets to create expanded, richer representation. Exploiting one of the similarity measures (such as widely used cosine coefficient), we now can apply any of successful clustering methods based on similarity such as HAC, K-means to cluster the enriched snippets. Our solution is simple, easy to implement, adaptable to multiple languages, and effectiveness in multiple applications.

Chapter 4: Matching and Ranking toward Online Contextual Advertising.

The problem of contextual advertising is based on the content to deliver ad messages, which normally consist of four parts: title, body, URL, and keywords, to the Web pages that users are surfing. It can therefore provide Internet users with information they are interested in and allow advertisers to reach their target customers in a non-intrusive way. In contextual advertising, one important observation is that the relevance between target Web pages and advertising messages is a significant factor to attract online users and customers. In order to suggest the "right ad messages, we need efficient and elegant contextual ad matching and ranking techniques. This chapter adapts the framework in Chapter 3 to the problem of matching, ranking for online advertising. By doing so, we show that our framework is adaptable and efficient in multiple applications.

Chapter 5: Feature-Word-Topic Model for Image Annotation and Retrieval

As high-resolution digital cameras become more affordable and widespread, the use of digital images is growing rapidly. At the same time, online photo-sharing websites (Flickr, Picasaweb, Photobucket, etc.) hosting hundreds of millions of pictures have quickly become an integral part of the Internet only after a couple of years. As a result, the need for better understanding of image data and multimedia data become increasingly important in order to make the Web more well-organized and accessible. Current commercial image retrieval systems are mostly based on text surrounding of images such as Google and Yahoo image search engines. Since they ignore visual representation of images, the search engines often return inappropriate images. Moreover, this approach cannot deal with images that are not accompanied with texts.

Content-based image retrieval, on the other hand, has become an active research topic over the last few years. While early systems were based on the query-by-example schema, which formalizes the task as search for best matches to example-images provided by users, the attention now moves to query-by-semantic schema in which queries are provided in natural language. This approach, however, needs a huge image database annotated with semantic labels. Due to the enormous number of photos taken every day, manual labeling becomes an extremely time-consuming and expensive task. As a result, automatic image annotation receives significant interest in image retrieval and multimedia mining.

Image annotation is a difficult task due to three problems namely *semantic gap*, *weakly labeling*, and *scalability*. The typical "semantic gap" problem is between low level features and higher level concepts. It means that extracting semantically meaningful concepts is hard when using only low level image features such as color or textures. The second problem, "weakly labeling", comes from the fact that exact mapping from keywords to image regions is usually unavailable. In other words, a label is given to an image without indications of which part of the image corresponds to that label. Since image annotation is served directly for image retrieval, "scalability" is also an important requirement and a problematic issue of image annotation.

This chapter presents a novel method for image annotation, which is based on feature-word and word-topic distributions. The main idea is to guess the scene settings or the story of the picture for image annotation. Suppose we have a picture of "grass field" to annotate, if we (human) see the picture, we first obtain the story of the picture such as "zebras on a grass fields with a blue sky above". Next, we can select "keywords" as labels based on it. Unfortunately, only based on "visual features", "sky" may be confused with "beach" learned from images with sea scene in the databases. If, somehow, we can guess scene settings of the picture, we can avoid such confusion since "zebras are not usually seen near a beach".

More specifically, we learn two models from the training dataset: 1) a model of feature-word distributions based on multiple instance learning and mixture hierarchies, which is like SML; 2) a model of word-topic distributions (topic model) estimated using probabilistic latent semantic analysis (pLSA). The models are concatenated to form feature-word-topic model for annotation, in which only words with highest values of feature-word distributions are used to infer latent topics of the image (based on word-topic distributions). The estimated topics are then exploited to re-rank words for annotation.

- The model is able to deal with the "weakly labeling" problem and optimize feature-word distributions. Moreover, since feature-word distributions for two different words can be estimated in a parallel manner, it is convenient to apply in real-world applications where the dataset is dynamically updated.
- Hidden topic analysis, which has shown the effectiveness in enriching semantic in text retrieval, is exploited to infer scene settings for image annotation. By doing so, we do not need to directly model word-to-word relationships and consider all possible word combinations, which could be very large, to obtain topic-consistent annotation. As a result, we can extend vocabulary while avoiding combinational explosion.
- Unlike previous generative models, the latent variable is not used to capture joint distributions among features and words, but among words only. The separation of topic modeling (via words only) and low-level image representation makes the annotation model more adaptable to different feature selection methods, or topic modeling.

Chapter 6: Cascade of Multi-level Multi-instance Classifiers for Image Annotation.

Image annotation is an important task to bridge the semantic gap in image retrieval. Although *image classification* and *object recognition* also assign meta data to images, the difference of image annotation from classification and recognition defines its typical challenging issues. In general, the number of labels (classes/objects) is usually larger in image annotation compared to classification and recognition. Because of the dominating number of negative examples, both the one-vs-one and one-vs-all schemes in multi-class supervised learning do not scale very well for image annotation. Unlike object recognition, image annotation is "weakly labeling, that is.a label is assigned to one image without indication of the region corresponding to that label. Moreover, scalability requirement prevents researchers investigating feature extraction for every label in image annotation. This, however, can be performed with a limited number of objects in object recognition. On the other hand, the variety of visual representations of objects suggests that we should not depend on one feature extraction method to work well with a large number of labels.

Motivated by the aforementioned issues, we propose a new learning method - a cascade of multilevel multi-instance classifiers (CMLMI) for image annotation. The idea behind our approach is that global features best describe the scene and common concepts such as "forest, building, mountain", while finer levels bring useful information to specific objects such as "tiger, cars, bear". Given an object, the cascade method ensures that we first detect the object's related scene, then focus on the "likely" scene to further recognize the object in that context. Formally, cascading means that learning classifiers at finer levels is dependent on classifiers at coarser levels (learning from coarse-to-fine). By so doing, when learning classifiers for specific objects at finer levels, we can ignore (negative) samples of non-related scenes, thus reduce training time. Since negative examples are those of the same scene without the considered object, there is more chance for us to separate the object from the background. For instance, since a "tiger" usually appears in a forest, the negative examples of forest background, which do not contain "tiger", helps recognize "negative" regions (or the forest regions).

論文審査結果の要旨

インターネットが普及した現在，情報検索はもっとも日常的に利用される情報処理技術であり，その高品質化は情報科学の学術及び情報産業の発展において非常に重要である．

本論文では，様々に表現された情報のもつ意味的なギャップの解消を自動的に行う理論を提案し，それを応用した情報検索の高品質化の手法を与えている．ユーザがWeb情報検索を行うとき，考える意図を表現する検索語を思い浮かべて入力し，その検索語を含むWebページのリストをシステムから得ることが一般的である．しかしながら，たとえば画像検索では，画像は検索語を文字として含むわけではなく，検索語を介したユーザの意図に対応した画像を出力するシステムを開発する必要がある．この，ユーザの意図(概念)，キーワード(言語表現)，データ画像(視覚表現)の 3 つの情報表現の間には本質的な意味的ギャップが存在する．このギャップを解消することが高品質の検索システムの構築には必要である．現在主流の画像検索システムでは，人手により画像へラベルや文書を付与することによるギャップの解消が行われているが，これを理論的に整備し，機械学習を用いて自動化することが重要な課題である．

著者は，トピックモデルという意味解析のための新規手法を提案し，それを用いてWeb検索とオンライン広告の高品質化を与え，さらにそれを一般化し，画像検索における自動ラベル付け問題の新規手法を開発し，システム実装による有効性を実証した．本論文は，これらの成果をとりまとめたものであり，全編 7 章からなる．

第 1 章は序論であり，本研究の背景と目的，過去の研究との関連を述べている．

第 2 章では，本論文で利用する意味解析の数理モデル化の一般的な 3 つの手法と，それらの数学的な理論基盤について解説している．

第 3 章では，言語の曖昧性に起因する意味的ギャップの解消に対して，トピックモデルの利用を提案している．これはトピック空間と言語空間の関連を抽出し，それを用いた意味解析を行うものであり，ベトナム語を例にして，言語に依存しない高品質Web検索システムの構築へ応用している．

第 4 章では，トピックモデルをオンライン広告問題へ応用し，提案手法がWebページを閲覧するユーザにとって必要性の高い広告をあたえ，有効な手法であることを実験により示している．

第 5 章では，画像検索において重要な画像注釈の自動付与の高品質化のために，トピックモデルを拡張し，画像から取り出した画像特性の作る統計量空間，言語空間，トピック空間の 3 つの空間の間の関係を利用したモデルの提案を行っている．これにより，画像と言語という 2 つの異なる意味表現の間の意味的ギャップ解消の改善に成功している．

第 6 章では，画像注釈問題の品質を改良するためにカスケード法を利用する新規な手法を提案し，第 5 章の手法と併せて，従来法に対する優位を実験的に検証している．

第 7 章は結論である．

以上要するに本論文は，高度情報検索に必要な意味的ギャップの解消のための数理的な新しいモデルの提唱とその理論展開をあたえ，重要な応用に対して提案手法の実装を行い，その優位性を示したものであり，システム情報科学の発展に寄与するところが少なくない．よって本論文は，博士（情報科学）の学位論文として合格と認める．