Cross-media Cross-genre Information Ranking Multi-media Information Networks

Tongtao Zhang Rensselaer Polytechnic Institute zhangt13@rpi.edu

Haibo Li Nuance lihaibo.c@gmail.com

Hongzhao Huang R.P.I. huangh9@rpi.edu

Heng Ji R.P.I. jih@rpi.edu

Min-Hsuan Tsai mtsai2@illinois.edu

Shen-Fu Tsai stsai8@illinois.edu

Thomas Huang University of Illinois at Urbana-Champaign huang@ifp.uiuc.edu

Abstract

Current web technology has brought us a scenario that information about a certain topic is widely dispersed in data from different domains and data modalities, such as texts and images from news and social media. Automatic extraction of the most informative and important multimedia summary (e.g. a ranked list of inter-connected texts and images) from massive amounts of cross-media and cross-genre data can significantly save users' time and effort that is consumed in browsing. In this paper, we propose a novel method to address this new task based on automatically constructed **M**ulti-media **I**nformation **Net**work**s** (**MiNets**) by incorporating cross-genre knowledge and inferring implicit similarity across texts and images. The facts from MiNets are exploited in a novel random walk-based algorithm to iteratively propagate ranking scores across multiple data modalities. Experimental results demonstrated the effectiveness of our MiNets-based approach and the power of cross-media cross-genre inference.

1 Introduction

Recent development on web technology – especially on fast connection and large-scale storage systems – has enabled social and news media to fulfill their jobs more efficiently in time and depth. However, such development also raises some problems such as overwhelming social media information and distracting news media contents. In emergent scenarios such as facing an incoming disaster (*e.g.*, Hurricane Irene in 2011 or Sandy in 2012), tweets and news are often repeatedly spread and forwarded in certain circles and contents are often overlapped by each other. However, browsing these messages and pages is almost unpleasant and inefficient. Therefore, an automatic summarization on piles of tweets and news is always necessary and welcomed, among which ranking is the most intuitive way to inform the users about the most informative content.

A passive solution is prompting the users to add more key words when typing the search query as most search engines do. However, without prior knowledge or due to the word limit, it is never trivial for the users to establish a satisfied ranking list for topics which attract more public attention. Recent changes on some Google Search have integrated image search and adopted some heterogenous content analysis, nevertheless, the connection between image and the keywords are still arbitrarily determined by the users, thus it is still far from optimal.

Active solutions which attempt to summarize information only focused on single data modalities. For example, Zanzotto et al. (2011) provided a comprehensive comparison about summarization methods for tweets. Zhao et al. (2011) developed a context-sensitive topical PageRank (Brin and Page, 1998) method to extract topical key phrase from Twitter as a way to summarize twitter content. As a new prospective, Feng and Lapata (2010) used LDA to annotate images, but this does not firmly integrate the information across different data types. Huang et al. (2012) presented a tweet ranking approach but only focused on single data modality (i.e., text).

Other conventional solutions towards analyzing the relationship or links between the instances have long been proposed and applied, such as PageRank (Brin and Page, 1998)and VisualRank (Jing and Baluja, 2008). The former is excessively used in heterogeneous networks (i.e., webpages and resources) but they are mainly based on linkage itself. VisualRank, which is based on PageRank, is a content-based linkage method but is confined with homogeneous networks.

Above all, our goal is to integrate cross-media inference and create the linkage among the information extracted from those heterogenous data. Our novel **M**ulti-media **I**nformation **Net**work**s** (**MiNets**) representation initializes our idea about a basic ontology of the ranking system.

The main contribution of this work is to fill in the domain gaps across different network genres and bridge them in a principled method. In this work, we manage to discover the hidden links or structures between the heterogeneous networks in different genres. We combine joint inference to resolve information conflicts across multi-genre

Proceedings of the 25th International Conference on Computational Linguistics, pages 74–81,
Dublin, Ireland, August 23-29 2014.

networks. We can also effectively measure, share and transfer complementary information and knowledge across multi-genre networks using structured correspondence.

The work is presented in sections as follows. We firstly introduce an overview of our system in Section 2. Detailed approaches in information extraction and constructing meta-information network are then followed in Section 3. Measurement across the multimedia information are proposed in Section 4 and 5. In Section 6 we demonstrate the results and performance gain.

2 Approach Overview

Within the context of an event where users generate a vast amount of multi-media messages in forms of tweets and images, we aim to provide a ranked subset of the most informative ones. Given a set of tweets $T = \{t_1, ..., t_n\}$, and a set of images $P = \{p_1, ..., p_m\}$ as input, our approach provides ordered lists of the most informative tweets or images (a.k.a objects) so that the informativeness of an object in position i is higher than or equal to that of an object in position $i + 1$. We consider the degree of informativeness of a certain object as the extent to which it provides valuable information to people who are involved in or tracking the event in question.

During emergent events, there are tight correlations between social media and web documents. Important information shared in social media tends to be posted in web documents. Therefore we also integrate information in a formal genre such as web documents to enhance the ranking quality of tweets and images. It consists of two main sub-tasks:

- ***Multimedia Information Network (MiNet) Construction***:
  Construct **MiNet** from cross-media and cross-genre information (i.e. tweets, images, sentences of web documents). Given a set of tweets and images on a specific topic as input, the formal genre web documents and images from the embedded URLs in those tweets are retrieved. Afterwards, a set of sentences and images are extracted from the web documents. Then we exploit advanced text Information Extraction and image Concept Extraction techniques to extract meta-information and construct the meta-information network. Together with three sets of heterogeneous input data, MiNet is constructed.

- ***MiNet-Based Information Ranking***: Rank the tweets and images. By extending and adapting Tri-HITS (Huang et al., 2012), we propose EN-Tri-HITs, an random walk-based propagation algorithm which iteratively propagate ranking scores for sentences, tweets, and images across **MiNet** to refine the tweet and image rankings.

3 Meta-information Network

When integrating information from different data modalities, meta-information network plays a pivotal role for representing interesting concepts and relations between them. We automatically construct the initial information networks using our state-of-the-art information extraction and image concept extraction techniques. A meta-information network is a heterogeneous network including a set of "information graphs" which is formally defined as: $G = \{G_i : G_i = (V_i, E_i)\}$, where V_i is the collection of concept nodes, and E_i is the collection of edges linking one concept to the other. An example is depicted in Figure 1. The meta-information network contains human knowledge pertaining to the target domain that could improve the performance of text process and image analysis. In this paper, we first construct meta-information networks separately from texts and images, and then fuse and enrich them through effective cross-media linking methods.

3.1 Information Extraction from Texts

Extracting salient types of facts for a meta-information network is challenging. In this paper we tackle this problem from two angles to balance the trade-off between quality and granularity/annotation cost. On one hand, to reveal deep semantics in meta-information network, we focus on achieving high-quality extraction for pre-defined fine-grained types such as those in NIST Automatic Content Extraction (ACE) [1]. For example, a "*Person/Individual*" node may include attributes such as "*Birth-Place*", and a "*Organization/Employee*" node may include attributes such as "*City-of-Headquarter*". These two nodes may be connected via a "*Employment/End-Position*" link.

We apply an Information Extraction (IE) system (Li et al., 2013) to extract entities, relations and events defined in ACE2005. There are 7 types of entities, 18 types of relations and 33 types of events. This system is based on a joint framework using structured perceptron with efficient beam-search and incorporating diverse lexical, syntactic, semantic and ontological features. We convert the IE output into the graph structured representation of meta-information network by mapping each entity as a node, and link entity nodes by semantic relations or events they are involved. For example, the relations between entities are naturally mapped to links in the meta-information network, such as the "*employment*" relation between "*Bill Read*" and "*Hurricane Center*". In addition, if an event

[1] http://www.itl.nist.gov/iad/894.01/tests/ace/

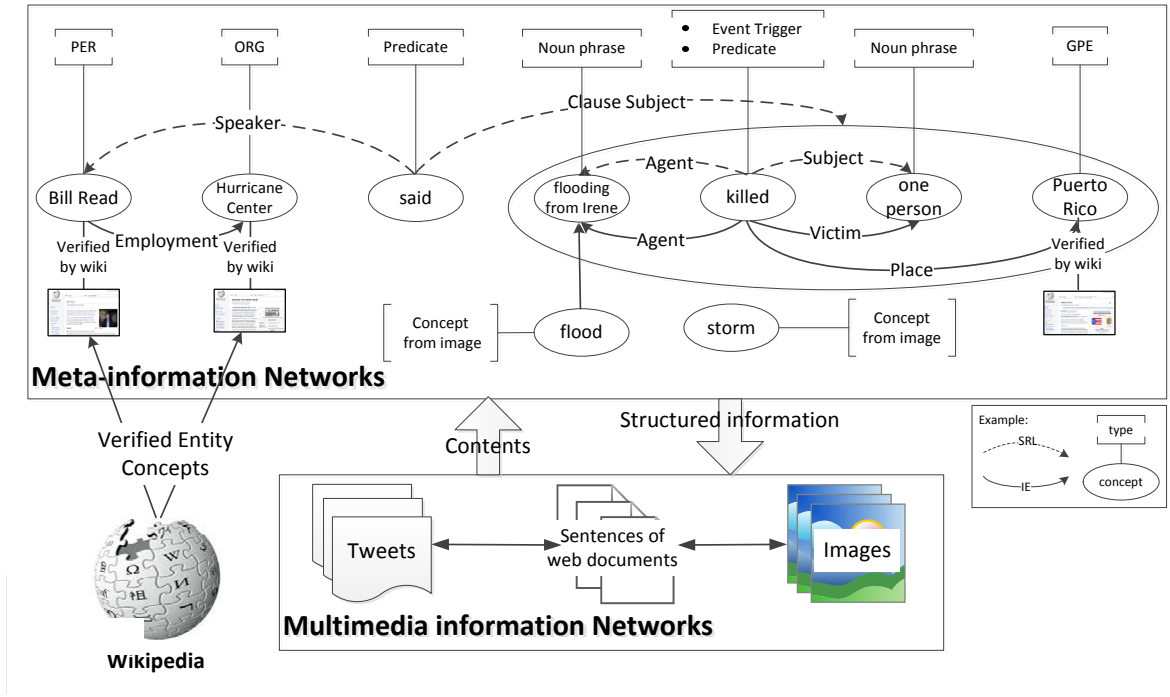

Figure 1: An example of meta-information network. Sentence: "Bill Read, Hurricane Center director, said that flooding from Irene killed at least one person in Puerto Rico"

argument is an entity, we also add an "*Event Argument*" link between the event trigger and the entity, such as the link between "*Irene*" and "*killed*".

On the other hand, in order to enrich the meta-information network, we extract more coarse-grained salient fact types based on Semantic Role Labeling (SRL) (Pradhan et al., 2008). For example, given the sentence "*In North Carolina, 10 counties are being evacuated.*", the "evacuation" event is not included in ACE. However, the SRL system can successfully detect the predicate ("evacuated") and its semantic roles ("10 counties" and "North Carolina"). These argument heads and predicates are added into the meta-information network as vertices, and edges are added between each predicate-argument pairs.

We merge entity mentions across tweets and web documents based on a cross-document entity clustering system described in (Chen and Ji, 2011). Moreover, for the same type of nodes from the SRL system, we also merge them by string matching across documents.

3.2 Concept Extraction from Images

We also developed a concept modeling approach by extending the similar framework in previous work (Tsai et al., 2012), Probabilistic Logical Tree (PLT), to extract semantic concepts from images. PLT integrates the logical and statistical inferences in a unifying framework where the existing primitive concepts are connected into a potentially unlimited vocabulary of high-level concepts by basic logical operations. In contrast, most existing image concept extraction algorithms either only learn a flat correlative concept structure, or a simple hierarchical structure without logical connections.

With an efficient statistical learning algorithm, the complex concepts in upper level of PLT are modeled upon some logically connected primitive concepts. This statistical learning approach is very flexible, where each concept in PLT can be modeled from distinctive feature spaces with the most suitable feature descriptors (e.g., visual features such as color and shape for scenery concepts).

For our case study on "Hurricane Irene" scenario, we apply this algorithm to extract the hierarchical concept trees with roots "flood" or "storm" from all the images in web documents whose URLs are contained in tweets.

The main problem is the classifications of the concepts such that it may be properly be placed onto an ontology. In order to enrich the hierarchy, we seek to classify these linkages through the use of the semi-structured and structured data that exists on Wikipedia. We use pattern matching to extract *is-a* relations from the first paragraphs of Wikipedia articles. For example, starting from our initial concept "Hurricane Irene", we can find its *is-a* relation with "Tropical Cyclone", and then climb up one more level to "Storm" where we can further mine lower concepts such as "Tornado" and "Snow Storm".

4 Multi-media Information Networks

A **M**ultimedia **In**formation **Net**work (*MINet*) is a structured collection made up of a set of multimedia documents (e.g., texts and images) and links between these documents. Each link corresponds to a specific relationship between nodes, such as hyperlinks between web documents or similarity links between tweets. In this paper, we construct our MINet based on two forms of contents from different domains: tweets, web documents (plain texts) and images.

4.1 Within-media Linking

4.1.1 Text-Text Similarity

Taking web document for example, we construct the meta-information network $G = \{G_i : G_i = (V_i, E_i)\}$ for all web documents D, in which each web document $d_i \in D$ corresponds to G_i. Given the meta-information network G, we compute the weight of each vertex $v_j \in V_i$ as $weight_{v_j} = \frac{nf(v_j,d)}{AVE(D)}$,

where $nf(v_j, d)$ is the mention number of node v_j appearing in a document d and $AVE(D)$ is the average number of mentions in a document d, which is defined as $AVE(D) = \frac{\sum_{d\in D} \text{concept mentions in } d}{|D|}$.

Similarly, we define the weight of each link $e_k \in E_i$ as $weight_{e_k} = \frac{nf(e_k,d)}{AVE(D)}$, where $nf(e_k, d)$ is the mention number of the node e_k in a document d and $AVE(D)$ is the average number of mentions in a document d, which is defined as $AVE(D) = \frac{\sum_{d\in D} \text{relation mentions in } d}{|D|}$.

If two edges share the same type and link nodes corresponding to the same tokens, we consider them as two mentions involved in a relation. Based on the weight of each concept mention and relation mention, we count their frequencies and transform them into vectors. Finally, we compute cosine similarity between every two vectors.

4.1.2 Image-Image Similarity

We extract Histogram of Oriented Gradients (HOG) features (Dalal and Triggs, 2005) from patches in images and apply Hierarchical Gaussianization (Zhou et al., 2009) to those HOG feature vectors. We learn a Gaussian mixture model (GMM) to obtain the statistics of the patches of an image by adapting the distribution of the extracted HOG features from these image patches and each image is represented by a super-vector. Based on the obtained image representation, the image-image similarity is simply a cosine similarity between two HG super-vectors.

4.2 Cross-media Linking

In order to obtain cross-media similarity, we propose a method based on transfer learning technique (Qi et al., 2012). Given a set of m points $[p_1, p_2, \ldots, p_m]$ in the source (image) domain $\mathcal{P}$, a set of n points $[t_1, t_2, \ldots, t_n]$ in the target (text) domain $\mathcal{T}$, and a set of N corresponding pairs $\mathcal{C} = \{(p_{a_i}, t_{b_i}\}_{i=1}^N$ in these two domains, we aim to find a cross-media similarity function:

$$G(p, t) = \ell((Up)^T(Vt)) = \ell(p^T St)), \tag{1}$$

where U and V are the linear embedding of $\mathcal{P}$ and $\mathcal{T}$, respectively. $S = U^T V$ is the *cross-domain similarity matrix* and $\ell(\theta) = \frac{1}{1+e^{-\theta}}$ is the logistic sigmoid function.

The key to S in Equation 1 is to solve the optimization problem blow:

$$\min_S \bar{\mathcal{L}}_s(S) + \lambda\bar{\mathcal{L}}_d(S) + \gamma\bar{\Omega}(S), \tag{2}$$

where $\bar{\mathcal{L}}_s(S) = \sum_{(x,y)\in\mathcal{C}} \log(1 + \exp(-p^T St))$, and $\bar{\Omega}(S) = \|S\|_*$ is the nuclear norm that is the surrogate of the matrix rank. Also, we have

$$\bar{\mathcal{L}}_d(S) = \frac{1}{2}\sum K_{\mathcal{P}}(p, p')d_{\mathcal{T}}(p, p') + \frac{1}{2}\sum K_{\mathcal{T}}(t, t')d_{\mathcal{P}}(t, t'),$$

where $K(\cdot, \cdot)$ is the similarity matrix among the points in a single domain and $d(\cdot, \cdot)$ defines the distance between two points due to the transfer.

Taking one step further, we have

$$\bar{\mathcal{L}}_d(S) = \text{tr}(L_{\mathcal{T}} Q_{\mathcal{T}}(S)^T K_{\mathcal{P}} Q_{\mathcal{P}}(S)) + \text{tr}(L_{\mathcal{X}} Q_{\mathcal{P}}(S)^T K_{\mathcal{T}} Q_{\mathcal{T}}(S)),$$

where $L_{\mathcal{P}}$ and $L_{\mathcal{T}}$ are the Laplacian matrices for $K_{\mathcal{P}}$ and $K_{\mathcal{T}}$, respectively.

To solve the optimization problem (2) with nuclear norm regularization we follow the proximal gradient method (Toh and Yun, 2010) with the following gradients:

$$\bigtriangledown\bar{\mathcal{L}}_s(S) = P(J_{\mathcal{C}} \circ H)P^T, \bigtriangledown\bar{\mathcal{L}}_d(S) = P((K_{\mathcal{P}} Q_{\mathcal{P}} L_{\mathcal{T}} + L_{\mathcal{P}} Q_{\mathcal{T}} K_{\mathcal{T}}) \circ H)P^T \tag{3}$$

$J_{\mathcal{C}}$ is an $m \times n$ matrix with its (i,j)-th entry 1 if $(p_i, t_j) \in \mathcal{C}$, otherwise 0. H is also an $m \times n$ matrix whose (i,j)-th entry where $H_{ij} = \ell'(p_i^T S t_j)$.

Hence we have

$$\bigtriangledown \bar{\mathcal{L}}(S) = P(\mathcal{G} \circ H)T^T, \tag{4}$$

where $\mathcal{G} = J_{\mathcal{C}} + \lambda K_{\mathcal{P}} Q_{\mathcal{P}} L_{\mathcal{T}} + \lambda L_{\mathcal{P}} Q_{\mathcal{T}} K_{\mathcal{T}}$. With the gradient in (4), one can solve the problem (2) using the proximal gradient method.

5 MiNet-Based Information Ranking: EN-Tri-HITs

5.1 Initializing Ranking Scores

Input: A set of tweets (T), and images (P) and web documents (W) on a given topic.
Output: Ranking scores (S_t) for T and (S_p) for P.

Use TextRank to compute initial ranking scores S_p^0 for P, S_t^0 for T and S_w^0 for W;
Construct multimedia information networks across P, T and W;
$k \leftarrow 0$, *diff* $\leftarrow 10e6$;
while $k <$ MaxIteration and *diff* $>$ MinThreshold **do**
    Use Eq. (5) (6) and (7) to compute S_p^{k+1}, S_t^{k+1} and S_w^{k+1};
    Normalize S_p^{k+1}, S_t^{k+1} and S_w^{k+1};
    diff $\leftarrow \max(\sum(|S_t^{k+1} - S_t^k|), \sum(|S_p^{k+1} - S_p^k|))$;
    $k \leftarrow k + 1$
end while

Algorithm 1: EN-Tri-HITS: Random walk on multimedia information networks

Graph-based ranking algorithms have been widely used to analyze relations between vertices in graphs. In this paper, we adapted PageRank (Brin and Page, 1998; Mihalcea and Tarau, 2004; Jing and Baluja, 2008) to compute initial ranking scores in tweet-only and image-only networks where edges between tweets or images are determined by their cosine similarity.

The ranking score is computed as follows:

$$S(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} S(V_j),$$

where V_i is a vertex with $S(V_i)$ as its ranking score; $In(V_i)$ and $Out(V_i)$ are the incoming edge set and outgoing edge set of V_i, respectively; w_{ij} is the weight for the edge between two vertices V_i and V_j. An edge links two vertices that represent text units when their cosine similarity of shared content exceeds or equals to a predefined threshold δ_t.

5.2 Random Walk on Multimedia Information Networks

We introduce a novel algorithm to incorporate both initial ranking scores and global evidence from multimedia information networks. It propagates ranking scores across MiNets iteratively. Our algorithm is a natural extension of Tri-HITS (Huang et al., 2012) based on the mutual reinforcement to boost linked objects.

By extending Tri-HITS, we develop enhanced Tri-HITS (EN-Tri-HITs) to handle multimedia information networks with three types of objects: Tweets (T), sentences of web documents (W) and images (P). EN-Tri-HITs is able to handle more complicated network structure with more links. Given the similarity matrices M^{tw} (between tweets and sentences of web documents), M^{wp} (between sentences of web documents and images) and M^{tp} (between tweets and images), and initial ranking scores of $S^0(p)$, $S^0(t)$ and $S^0(w)$, we aim to refine the initial ranking scores and obtain the final ranking scores $S(w)$, $S(t)$ and $S(p)$. Starting from images $S(p)$, the update process considers both the initial score $S^0(p)$ and the propagation from connected tweets $S(t)$ and web documents $S(w)$, which can be expressed as:

$$\hat{S}_w(p_j) = \sum_{i \in W} m_{ij}^{wp} S(w_i), \hat{S}_t(p_j) = \sum_{k \in T} m_{kj}^{tp} S(t_k),$$

$$S(p_j) = (1 - \lambda_{wp} - \lambda_{tp}) S^0(p_j) + \lambda_{wp} \frac{\hat{S}_w(p_j)}{\sum_j \hat{S}_w(p_j)} + \lambda_{tp} \frac{\hat{S}_t(p_j)}{\sum_j \hat{S}_t(p_j)}, \tag{5}$$

Set ID	Tweets	Web Doc (Sentences)	Images
1	1171	41(1272)	183
2	1116	47(1634)	265
3	1184	69(1639)	346
All	3471	157(4545)	794

Table 1: Data Statistics: Numbers of each item in the dataset.

	word	word +IE	word +SRL	word +IE+SRL
I+W	0.545	0.539	0.521	0.583
I+T	0.422	0.436	0.407	0.489
I+W+T	0.526	0.513	0.492	0.541

Table 2: NDCG@5 of Images. The image ranking baseline performance is 0.421. **I** stands for Image; **W** Web Documents; **T** Tweets

where $\lambda_{wp}, \lambda_{tp} \in [0,1]$ ($\lambda_{wp} + \lambda_{tp} \leq 1$) are the parameters to balance between initial and propagated ranking scores. Similar to Tri-HITS, EN-Tri-HITS normalizes the propagated ranking scores $\hat{S}_w(p_i)$ and $\hat{S}_t(p_i)$.

Similarly, we define the propagations from images and web documents to tweets as follows:

$$\hat{S}_p(t_k) = \sum_{i \in P} m_{ik}^{pt} S(p_i), \hat{S}_w(t_k) = \sum_{j \in W} m_{jk}^{wt} S(w_j),$$
$$S(t_k) = (1 - \lambda_{wt} - \lambda_{pt}) S^0(t_k) + \lambda_{wt} \frac{\hat{S}_p(t_k)}{\sum_k \hat{S}_p(t_k)} + \lambda_{pt} \frac{\hat{S}_w(t_k)}{\sum_k \hat{S}_w(t_k)}, \quad (6)$$

where M^{pt} is the transpose of M^{tp}, λ_{pt} and λ_{wt} are parameters to balance between initial and propagated ranking scores.

Each sentence of web documents $S(w_j)$ may be influenced by the propagation from both tweets and images:

$$\hat{S}_t(w_i) = \sum_{k \in T} m_{ki}^{tw} S(t_k), \hat{S}_p(w_i) = \sum_{i \in P} m_{ji}^{pw} S(p_j),$$
$$S(w_i) = (1 - \lambda_{tw} - \lambda_{pw}) S^0(w_i) + \lambda_{tw} \frac{\hat{S}_t(w_i)}{\sum_i \hat{S}_t(w_i)} + \lambda_{pw} \frac{\hat{S}_p(w_i)}{\sum_i \hat{S}_p(w_i)}, \quad (7)$$

where M^{pw} is the transpose of M^{wp}, λ_{tw} and λ_{pw} are parameters to balance between initial and propagated ranking scores.

Algorithm 1 summarizes En-Tri-HITS.

6 Experiments

6.1 Data and Scoring Metric

Currently there are no information ranking related benchmark data sets publicly available, therefore we build our own data set and network ontology.

We crawled 3471 tweets during a three-hour period and extracted key phrases from these tweets, then we use the key phrases as image search queries. The image search queries are submitted to Bing Image Search API and we take the top 10 images for each query. We extract a 512-d GIST feature from each image for meta information training. For image similarity metrics, we resize images to a maximum of 240×240 and segmented into patches with three different sizes (16, 25 and 31) by a 6-pixel step size. A 128-d Histogram of Oriented Gradients (HOG) feature is extracted from each patch and followed by a PCA dimension reduction to 80-d. The size of dimension of the final feature vector for each image is 42,496.

We create the ground truth based on human assessment of informativeness on a 5-star likert scale, with grade 5 as the most informative and 1 as the least informative. Table 1 presents an overview on our data sets. We conduct 3-fold cross-validation for our experiments.

To evaluate tweet ranking, we use *nDCG* as our evaluation metric (Järvelin and Kekäläinen, 2002), which considers both the informativeness and the position of a tweet:

$$nDCG(\Phi, k) = \frac{1}{|\Phi|} \sum_{i=1}^{|\Phi|} \frac{DCG_{ik}}{IDCG_{ik}}, DCG_{ik} = \sum_{j=1}^{k} \frac{2^{rel_{ij}} - 1}{\log(1+j)},$$

where Φ is the set of documents in the test set, with each document corresponding to an hour of tweets in our case, rel_{ij} is the human-annotated label for the tweet j in the document i, and $IDCG_{ik}$ is the DCG score of the ideal ranking. The average *nDCG* score for the top k tweets is: $Avg@k = \sum_{i=1}^{k} nDCG(\Phi, i)/k$. To favor diversity of top ranked tweets, redundant tweets are penalized to lower down the final score.

6.2 Impact of Cross-media Inference

Table 2 and Figure 2 present the image ranking results. The results indicate that methods integrating heterogeneous networks outperform the baseline of image ranking (0.421). When web documents are aligned with images (row 1), the ranking quality improves significantly, proving that web documents can help detect informative images by adding support from text media of formal genre. However, the text media of informal genre, such as tweets, almost cannot help improve the ranking performance.

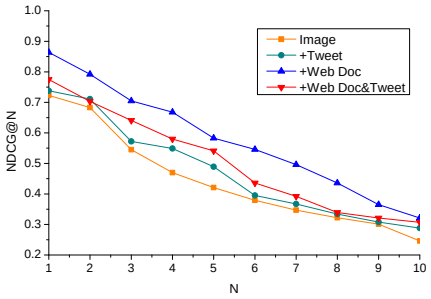

Figure 2: NDCG@n score of Images with Various n

	word	word +IE	word +SRL	word +IE+SRL
T	0.675	0.691	0.697	0.700
T+W	0.766	0.771	0.757	0.809
T+I	0.675	0.691	0.667	0.700
T+W+I	0.722	0.771	0.757	0.809

Table 3: NDCG@5 of Tweets

6.3 Impact of Cross-genre Inference

Methods that integrate heterogeneous networks after filtering, outperform the baseline TextRank, as shown in Table 3. When tweets are aligned with web documents, the ranking quality improves significantly, proving that web documents can help infer informative tweets by adding support from a formal genre. The fact that tweets with low initial ranking scores are aligned with web documents helps promote their ranking positions. For example, the ranking of the tweet "*Hurricane Irene: City by City Forecasts http://t.co/x1t122A*" is improved compared to TextRank, benefitting from the fact that 10 retrieved web documents are about this topic.

6.4 Remaining Error Analysis

Enhanced Tri-HITS shows encouraging improvements in ranking quality with respect to a state-of-the-art model such as TextRank. However, there are still some issues to be addressed for further improvements.

(i) Long tweets preferred. We tracked tweets containing the keywords "Hurricane" and "Irene". Using such a query might also return tweets that are not related to the event being followed. This may occur either because the terms are ambiguous, or because of spam being injected into trending conversations to make it visible. For example, the tweet "*Hurricane Kitty: http://t.co/cdIexE3*" is an advertisement, which is not topically related to Irene.

(ii) Deep semantic analysis of the content, especially for images. We rely on distinct terms to refer to the same concept. More extensive semantic analyses of text can help identify those terms, possibly enhancing the propagation process. For example, we can explore existing text dictionaries such as WordNet (Miller, 1995) to mine synonym/hypernym/hyponym relations, and Brown clusters (Brown et al., 1992) to mine other types of relations in order to enrich the concepts extracted from images.

7 Conclusion and Future Work

In this paper, we propose a comprehensive information ranking approach which facilitates measurement on cross-media/cross-genre informativeness based on a novel multi-media information network representation **MiNet**. We establish links via information extraction method from text and images and verification with Wikipedia. In addition, we propose similarity measurement on intra-media and cross-media using transfer learning techniques. We also introduce a novel En-Tri-Hits algorithm to evaluate the ranking scores across **MiNet**. Experiments have demonstrated that our cross-media/cross-genre ranking method is able to significantly boost the performance of multi-media tweet ranking. In the future, we aim to focus on enhancing the quality of concept extraction by exploiting cross-media inference that goes beyond simple fusion.

Acknowledgement

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), U.S. NSF CAREER Award under Grant IIS-0953149, U.S. DARPA Award No. FA8750-13-2-0041 in the ?Deep Exploration and Filtering of Text? (DEFT) Program, IBM Faculty award and RPI faculty start-up grant. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The

References

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Zheng Chen and Heng Ji. 2011. Collaborative ranking: A case study on entity linking. In *Proc. EMNLP2011*.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893.

Yansong Feng and Mirella Lapata. 2010. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 831–839, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hongzhao Huang, Arkaitz Zubiaga, Heng Ji, Hongbo Deng, Dong Wang, Hieu Le, Tarek Abdelzaher, Jiawei Han, Alice Leung, John Hancock, and Clare Voss. 2012. Tweet ranking based on heterogeneous networks. In *Proc. COLING 2012*, pages 1239–1256, Mumbai, India. The COLING 2012 Organizing Committee.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October.

Yushi Jing and Shumeet Baluja. 2008. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1877–1890.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proc. ACL2013*, pages 73–82.

R. Mihalcea and P. Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of EMNLP*, volume 4. Barcelona: ACL.

George A. Miller. 1995. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Sameer Pradhan, Wayne Ward, and James H. Martin. 2008. Towards robust semantic role labeling. In *Computational Linguistics Special Issue on Semantic Role Labeling*, volume 34, pages 289–310.

Guo-Jun Qi, Charu C. Aggarwal, and Thomas S. Huang. 2012. Transfer learning of distance metrics by cross-domain metric sampling across heterogeneous spaces. In *SDM*, pages 528–539.

Kim-Chuan Toh and Sangwoon Yun. 2010. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*.

Shen-Fu Tsai, Henry Hao Tang, Feng Tang, and Thomas S. Huang. 2012. Ontological inference framework with joint ontology construction and learning for image understanding. In *IEEE International Conference on Multimedia and Expo (ICME) 2012*.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Kostas Tsioutsiouliklis. 2011. Linguistic redundancy in twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 659–669, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wayne X. Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee P. Lim, and Xiaoming Li. 2011. Topical keyphrase extraction from Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 379–388, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xi Zhou, Na Cui, Zhen Li, Feng Liang, and Thomas S. Huang. 2009. Hierarchical gaussianization for image classification. In *ICCV*, pages 1971–1977.