Information Diffusion in Twitter across Different Language Groups

Chulki Lee
School of Information
UC Berkeley
Berkeley, CA 94720
chulkilee@ischool.berkeley.edu

Eungchan Kim
Graduate School of Convergence Science and Technology
Seoul National University
Seoul, Korea
yangpa15@snu.ac.kr

ABSTRACT
As Twitter has become a [communication platform], information diffusion in Twitter has received much attention. To discover how information is diffused, various empirical quantitative studies have done. [3] [4] These studies assume one large information network which miss the fact that popular websites are used by people over the world. In this sense, this paper examines how language barriers impacts on information diffusion. We collect [TODO] and [TODO]..... The result show that [TODO].

Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: Miscellaneous; J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms
Human Factors

Keywords
Information diffusion

1. INTRODUCTION

1.1 Research questions
Here are possible research questions. We will choose some of them.

- How do information diffuse across countries where use different languages on Twitter? between English and Korean
- Are there the certain number of gateways for information diffusion between different languages?
- What causes the difference of information diffusion between different languages?
- Is the direction of information diffusion from English to other languages dominant?
- Do the time differences of diffusing information have a significant difference according to Topics ?
- Is language difference really a barrier within a virtual community especially in the blogosphere, wikis and social networking sites?
- Who diffuses information across networks using different languages?
- How does language difference in the Internet affect the directionality of information flow?

2. BACKGROUND

2.1 Why Twitter?[TODO]
In 2010, according to press release by Semiocast, half of tweets were not in English and Japanese took up 14% of tweets. [5].

earthquake detection predicting popularity

2.2 Models for Information Diffusion
To explain general information diffusion, several models are developed: Threshold models, Cascade models, Epidemic models. Here are several reviews.

2.2.1 Social network threshold model of the diffusion of innovations [6]
Key Findings: The early adopters have more sources of external influence. This characteristics of the early adopters can be applied to that of a gatekeeper who controls the flow of information from one languages(English) to another language(Japan). This is also supported by one of two possible external sources, cosmopolitan action, a tendency of orienting to the world outside of his/her local social system and linking his/her local one to the larger environment by providing links to outside information(Gouldner, 1957, 1958: Davis, 1961). Moreover, according to the author's empirical analysis, various adopters can be categories by the 4x4 matrix with both personal network thresholds (time-of-adoptions) and social network ties.

History of Network approach to diffusion research

1. In-degree distribution (Rogers, 1962): the number of times an individual was nominated as a network partner.
2. Structural approach: weak tie (Granovetter, 1973, 1982)
3. Structural equivalence: the degree of equality in network position (Burt, 1980, 1987), Centrality, density and Reciprocity (Rice, 1994; Valente, 1995)
4. Threshold model: an individual engages in a behavior based on the proportion of people in the social system already engages in the behavior (Granovetter, 1978)

Social network threshold model of the diffusion of innovations: based on the Ryan and Gross(1943) adopter categories

- *early adopters* : individuals whose time-of-adoption is greater than one standard deviation earlier than the average time-of-adoption
- *early majority* and *late majority* : individuals whose time-of-adoption is bounded by one standard deviation earlier and later than the average
- *laggards* : individuals who adopted later than one standard deviation from the mean

2.2.2 Networks, Crowds, and Markets. Reasoning about a Highly Connected World [2]

Key Findings: The author tries to describe information cascading phenomenon from both more individual level perspective and fine structure level of the network perspective. The former is simply demonstrated by the Bayes' Theorem, and quite well applied to real-world situation. The latter models fits our interests, an information cascading model within a specific language-use group.

Information Cascade: the level of assuming amorphous population of individuals, and looking at effects in aggregate

- Why imitating the behavior of others can be beneficial:
  1. informational effects: the choices made by others can provide indirect information about what they know
  2. direct-benefit effects: payoffs that arise from using compatible technologies instead of incompatible ones.
- Prior decisions made by others can impact on posterior decision. For instance, even though each person has his/her own private information, when s/he observes other people's decisions which are different from their own information, it is high possible of him/her to follow initial majorities' opinion. This model can be simply demonstrated by the Bayes' Rule.

Cascading Behavior in Networks: the level of fine structure of the network as a graph, and looking at how individuals are influenced by their particular network neighbors.

- assumption: many of our interaction with the rest of the world happen at a local, rather than a global, level
- The diffusion of innovations (Rogers): The success of an innovation also depends on its:
  - *complexity* for people to understand and implement
  - *observability*, so that people can become aware that others are using it
  - *trialability*, so that people can mitigate its risks by adopting it gradually and incrementally
  - overall *compatibility* with with the social system that it is entering. The principle of homophily
- The principle of homophily: sometimes act as a barrier to diffusion: since people tend to interact with others who are like themselves, while new innovations tend to arrive from "outside" the system, it can be difficult for these innovations to make their way into a tightly-knit social community.

2.3 Empirical studies

Many quantitative and empirical studies - distribution of participation [3] - differences across topics [4] - whom to speak with: homophily.. [8] [1]

However, existing studies missed one of the most important emerging factor - global use of the Internet -> different languages

2.3.1 A measurement-driven analysis of information propagation in the flickr social network

Key Findings: The authors' research questions, how widely and how quickly information propagates in Flickr sphere, are similar to our concerns on Twitter. We presumably segment Twitter users, based on languages they use. We more focus on analyzing information propagation within each group and between groups. To do so, we are able to use the authors' empirical research methods for collecting data and analyzing data. They used a random snowballing sampling by using Flickr API, and took daily snapshots of social graph network so that they were able to use heuristic method for identifying social cascade phenomenon.

3. METHOD

3.1 Topics

To examine information diffusion across different groups, we have to choose topics covered by all groups. For example, if we choose language or culture specific topics, the difference of attention might influence more than the language difference between groups. Although it is clear that we cannot eliminate such factors, choosing various topics deliberately could decrease the influence.

To cover various topics, we considered [TODO]. Considering them, we chose following topics.

- Topic A

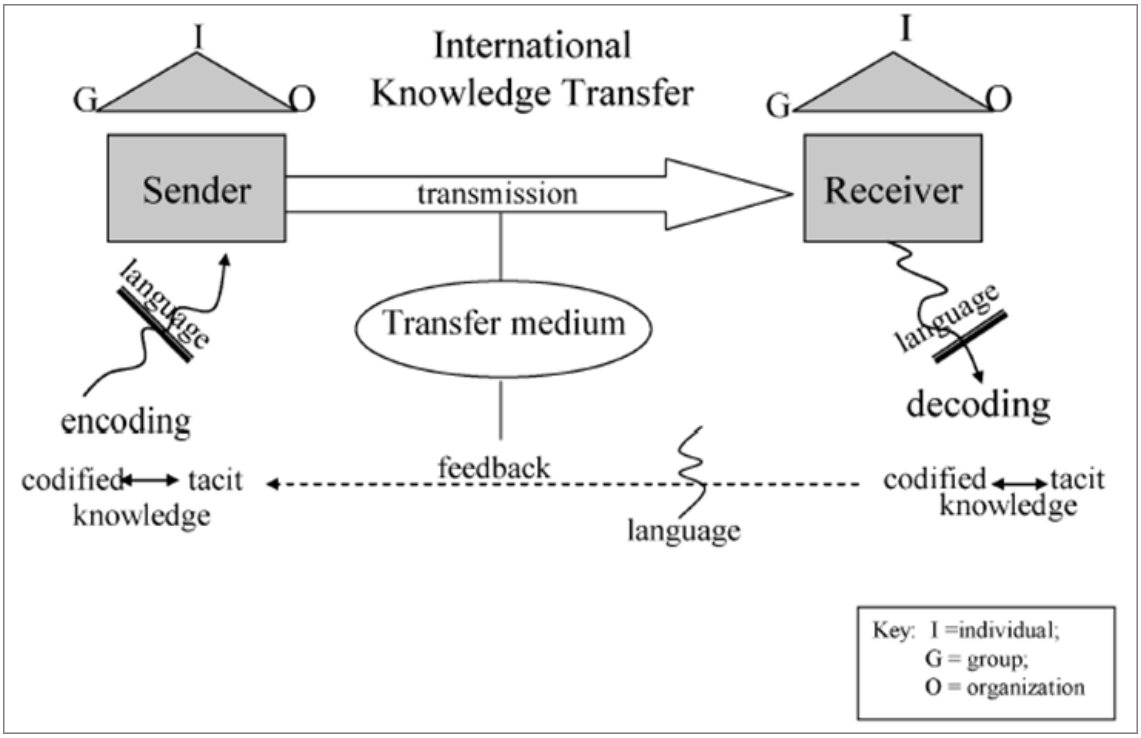

Figure 1: International knowledge transfer [7]

Table 1: Search keywords for topics

Topic	Search keywords in English
Topic A	keywordA, keywordB

3.2 Groups

We chose English and Japanese. First, they are the fist and the second the second most used language in Twitter [5]. Second, there are less bilingual users speaking them than other languages, such as English and Spanish. Third, people using the languages are far from each other geographically so that there are less opportunities to interact at offline. These characteristics make [TODO language barrier?]

3.3 Data collection

Twitter provides an Application Programming Interface (API) to access various types of data. We built a python Twitter crawler using the API[1]. ***Currently we are testing the crawler and choosing topics to crawl.***

3.3.1 Topics

To collect tweets on specific topics, we used keyword-based search using *statuses/filter* method. Since people use different words, phrases or hashtags to mention a single event, we used several search keywords for each topic. Following tables shows used keywords for topics.

3.3.2 Tweets and profiles

[1]Chulki Lee is in another project group, *Influential Tweeple*. He built a crawler for both teams

Table 2: Overview of collected tweets

	English	Japanese
# of tweets	[TODO]	[TODO]
# of unique users who tweeted	[TODO]	[TODO]
# of retweets	[TODO]	[TODO]

We collected all tweets and profiles of users who tweeted starting on [TODO] until [TODO]. In addition, to inspect network structure we collected past [TODO] tweets and [TODO] profiles of the users.

3.3.3 Languages

To determine a language of each user, we guess the language from all tweets from the user using *guess-language*. Twitter API gives a language of a user, which indicates interface language, but [TODO multilingual users...]

3.4 Network topology analysis

First we need to know how they are connected to each other. [TODO]

3.5 Temporal analysis

To understand how information is diffused across groups, we need to look into various aspects of information diffusion. In addition, we need to see both short-period and long-period trends.

4. RESULTS

During the study period, [TODO] tweets are collected.

4.1 Network topology

We will provide following graphs.

- in- and out-degrees between groups
- Degree of separation [3]
- The average time differences between a user and r-friends [3]

4.2 Temporal data

We will provide following graphs.

- Volumes of tweets over time
- Number of participants over time
- Time lag between a retweet and the original tweets over time [3]
- Cumulative numbers of tweets and users over time [3]
- Cumulative fraction: # of active periods / topic and Duration of active period [3]
- Influence curves [4]

5. DISCUSSION & CONCLUSION

5.1 Information diffusion models

We will discuss our findings with several information diffusion models.

5.2 Comparison with other empirical studies

We will compare our findings with other empirical studies which did not consider language barriers.

5.3 Limitation

First, we cannot control out influence of other factors. For example, the differences may come from different characteristics of populations, not from language barriers.

6. REFERENCES

[1] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 721–730, New York, NY, USA, 2009. ACM.

[2] E. David and K. Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[3] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[4] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 695–704, New York, NY, USA, 2011. ACM.

[5] Semiocast. Half of messages on twitter are not in english: Japanese is the second most used language, Feb. 2010.

[6] V. Thomas W. Social network thresholds in the diffusion of innovations. *Social Networks*, 18(1):69–89, Jan. 1996.

[7] D. Welch and L. Welch. The importance of language in international knowledge transfer. *Management International Review*, 48(3):339–360, 2008. 10.1007/s11575-008-0019-7.

[8] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 705–714, New York, NY, USA, 2011. ACM.