The Electronic Library

Comparative analysis of book tags: a cross-lingual perspective

Chao Lu Chengzhi Zhang Daqing He

Received 13 March 2015
Revised 19 May 2015
28 July 2015
Accepted 16 September 2015

Comparative analysis of book tags: a cross-lingual perspective

Chao Lu and Chengzhi Zhang
Department of Information Management, Nanjing University of Science and Technology, Nanjing, China, and
Daqing He
School of Information Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Abstract

Purpose – In the era of social media, users all over the world annotate books with social tags to express their preferences and interests. The purpose of this paper is to explore different tagging behaviours by analysing the book tags in different languages.

Design/methodology/approach – This investigation collected nearly 56,000 tags of 1,200 books from one Chinese and two English online bookmarking systems; it combined content analysis and machine-processing methods to evaluate the similarities and differences between different tagging systems from a cross-lingual perspective. Jaccard's coefficient was adopted to evaluate the similarity level.

Findings – The results show that the similarity between mono-lingual tags of the same books is higher than that of cross-lingual tags in different systems and the similarity between tags of books written for specialties is higher than that of books written for the general public.

Research limitations/implications – Those who have more in common annotate books with more similar tags. The similarity between users in tagging systems determines the similarity of the tag sets.

Practical implications – The results and conclusion of this study will benefit users' cross-lingual information retrieval and cross-lingual book recommendation for online bookmarking systems.

Originality/value – This study may be one of the first to compare cross-lingual tags. Its methodology can be applied to tag comparison between any two languages. The insights of this study will help develop cross-lingual tagging systems and improve information retrieval.

Keywords Book annotation, Common space, Social tag, User tagging behaviour

Paper type Research paper

1. Introduction

Social tagging (social annotation) has gained extensive attention with the development of the internet and social media sites, such as Delicious (https://delicious.com), Flickr (www.flickr.com) and LibraryThing (www.librarything.com). On these social media sites, users collect or share pictures, videos, music, URLs, books and other online

This work was supported in part by National Social Science Fund Project (No.14BTQ033), National Natural Science Foundation of China (No. 70903032), Project of the Education Ministry's Humanities and Social Science (No. 13YJA870020), Major Projects of National Social Science Fund (13&ZD174) and the Fundamental Research Funds for the Central Universities (No. 30915011323). Thanks are due to Prof. Heting Chu, Dr Kun Huang and Dr Bikun Chen for their valuable suggestions on this study.

The Electronic Library
Vol. 34 No. 4, 2016
pp. 666-682
© Emerald Group Publishing Limited
0264-0473
DOI 10.1108/EL-03-2015-0042

resources using social tags (Strohmaier *et al.*, 2012). Social tags help users store, retrieve and share online resources more conveniently than ever. Hence, social tagging has become one of the major information organization tools for online resources. Various social annotation systems have gained attention from libraries and scholars in related fields. A number of libraries are using tag clouds and tag recommendation technologies to support traditional book retrieval and to improve collection retrieval, and to develop related library services. Theoretically, many studies attempted to compare social tags with controlled vocabularies and explore the differences and similarities. Rolla (2009) concluded that tags and Library of Congress Subject Headings are very different. Several studies reached similar conclusions to Rolla's study (Lee and Schleyer, 2012; Lu *et al.*, 2010; Wu *et al.*, 2013). While subject headings focus on the subject of a resource, tags cover other aspects of a resource (Tennis, 2006), such as the users' opinion of it. Only a few studies have investigated cross-platform tags of the same book resource, such as Jung (2012), and no studies have investigated cross-lingual tags of the same book. Assume that a user is in a foreign language environment and he wants to find a book using tags in his native language. Will he succeed? In other words, will the tags in one language help him find books in another language and improve the cross-lingual recommendation? The answers will help us better understand social tags at the cross-lingual level and direct the next stage of investigation – using cross-lingual tags to improve cross-lingual book retrieval and book recommendations. So far, however, no such studies have been done.

The cross-lingual tags function in two ways. Some social tagging systems serve users from different countries. For example, in LibraryThing, a book can be annotated with tags in English, Japanese and other languages simultaneously; however, most of the tags are in English. Other social tagging systems provide similar services, but only for native users. For example, Douban Reading (http://book.douban.com), a Chinese book tagging system, provides book tagging services mainly for Chinese users. Many books in LibraryThing and Douban Reading have English as well as Chinese tags. This study aims to compare tags of the same books to find out whether English users and Chinese users think alike and annotate the same books with similar tags and, if not, to identify what differences may exist. After this investigation, something new in cross-lingual tagging features may be found that has never been observed before. The features would be used to direct the construction of cross-lingual tagging systems and improve cross-lingual tagging services, such as cross-lingual tag searching (Elhussein and Nakata, 2010; Jung, 2010) and recommendations. Also, the results could help users search books in a more convenient way.

To fulfil the aims of this study, three websites Douban Reading, LibraryThing and Amazon (www.amazon.com) were chosen from which English and Chinese tag data were collected for the same books. The third website, Amazon, is introduced to provide a reference when comparing the results. Douban Reading and LibraryThing both serve readers and provide book tagging, while Amazon, which is famous for online bookselling, provides book tagging that allows buyers to express their opinions of books as well. Content analysis and machine processing are combined to evaluate the similarities and differences between different tagging systems from a cross-lingual perspective.

This article is organized as follows: Section 2 reviews related articles concerning social tagging analysis, tagging comparison and cross-lingual studies; Section 3 introduces the research design of this investigation on research questions, data and methodology; Section 4 presents the findings, followed by the discussion part; finally, Section 5 concludes this research.

2. Literature review

Once Web 2.0 and social media sites became popular, they attracted the research interest of many scholars and information professionals. Some studies discussed social tags with classic topics, such as information organization (Golub *et al.*, 2014; Tennis, 2006; Wetzker *et al.*, 2008; Yi and Chan, 2009) and information retrieval (Carman *et al.*, 2008; DeZelar-Tiedman, 2011; Gelernter, 2007; Guerra and LaPlante, 2011; Li *et al.*, 2008; Ruiz and Chin, 2010); some investigated the types of tags (Golder and Huberman, 2006; Gupta *et al.*, 2010; Thomas *et al.*, 2009); some focused on tagging behaviour and the motivation behind it; some studies paid attention to the relationship between book tags and controlled vocabularies, such as subject headings or Dublin Core (Catarino and Baptista, 2010), and the application of social tags in library service (DeZelar-Tiedman, 2011; Gelernter, 2007; Kakali and Papatheodorou, 2010; Lu *et al.*, 2010; Peters, 2009; Thomas *et al.*, 2009).

2.1 Social tagging analysis

Many studies have focused on tagging behaviours (Chen and Ke, 2013; De Meo *et al.*, 2013; Farooq *et al.*, 2007; Golbeck *et al.*, 2011; Lin and Chen, 2012; Ruiz and Chin, 2010; Santos-Neto *et al.*, 2009; Vuorikari *et al.*, 2007; Wan *et al.*, 2013) and the motivation behind them (Ames and Naaman, 2007; Elhussein and Nakata, 2012; Gupta *et al.*, 2010; Strohmaier *et al.*, 2010).

Numerous tagging behaviour studies examined the general tagging distribution and tagging vocabularies and used the results to describe tagging behaviour. Kipp and Campbell (2006) investigated tags from the Delicious site and found that tag frequency follows the power-law distribution. Other researchers also observed a similar pattern (Chen and Ke, 2013; Farooq *et al.*, 2007; Golder and Huberman, 2006; Ke and Chen, 2012; Yi and Chan, 2009). The linguistic part of social tags was also studied. Kipp and Campbell (2006) identified problems with tags on the Delicious site, such as acronyms, synonyms and spelling variations. Spiteri (2013) observed that users tagged more nouns, such as abbreviations, acronyms and homographs, than other grammatical forms, such as verbs. Another study on Connotea (www.connotea.org) (Heckner *et al.*, 2008) claimed that 72 per cent of single-word user tags were nouns and none were verbs. They also found many acronyms in the tag data. Farooq *et al.* (2007) studied over two years of tagging data from CiteULike from six aspects – tag growth, tag reuse, tag non-obviousness, tag discrimination, tag frequency and tag patterns – to investigate user behaviour and patterns to support and complement existing social bookmarking system design.

Several scholars have studied tagging motivation. Strohmaier *et al.* (2010) measured the ESP game data set (www.cs.cmu.edu/~biglou/resources/) and found that users show different motivations within a tagging system, as well as across different ones, and users agree more with tags describing resources than those categorizing resources. Ames and Naaman (2007) built a taxonomy of tagging motivations for Flickr and ZoneTag in two dimensions – function (organization and communication) and sociality (self, friends/families and public). They analysed tagging data on Flickr and ZoneTag and deployed semi-structured interviews with 13 participants. The results suggested that more tags are used to organize resources for users and the public and to communicate with friends and families. Gupta *et al.* (2010) identified ten kinds of user tagging motivations in their paper: future retrieval, contribution and sharing, attract attention, play and competition, self-presentation, opinion expression, task organization, social signalling, money and technological ease.

2.2 Social tags comparative studies

A number of studies have focused on comparing social tags with other kinds of indexing or controlled vocabularies. Tennis (2006) used a framework analysis and compared social tagging with subject cataloguing in detail. He pointed out that indexing seems incipient in the new technological environment, while social tagging is useful for identifying the explicit links to intertextuality, authorship and task; he also claimed that indexing is under-nourished and falls behind because of the unceasing innovation within the online environment, but social tagging works well for online resources. Lee and Schleyer (2012) thought differently. They collected 76,968 distinct tags and 21,129 distinct Medical Subject Heading (MeSH) terms from more than 200,000 papers in CiteULike. They used Jaccard's coefficient, coverage ratio and other measures to investigate the data paper-by-paper. The results suggested that the similarity is quite low – only 2.12 to 3.3 per cent. They believed that social tagging is totally different from MeSH, and there is no substitute for controlled indexing. After comparing tags in LibraryThing with subject headings in the Library of Congress, Rolla (2009) concluded that a better means of indexing is to combine social tagging with subject headings. Further studies adopted a similar method and compared social tags in LibraryThing with subject headings in the Library of Congress to control social tagging (Golub *et al.*, 2014; Yi and Chan, 2009) and complement traditional indexing (Bartley, 2009; DeZelar-Tiedman, 2011; Golub *et al.*, 2014; Lawson, 2009; Lu *et al.*, 2010; Thomas *et al.*, 2009; Wu *et al.*, 2013). Some studies on images and comparable materials made similar comparisons between tags and controlled vocabularies (Golbeck *et al.*, 2011; Petek, 2012; Rorissa, 2010).

2.3 Cross-lingual tagging analysis

In cross-lingual tagging analysis, there are only rare studies on book annotations and a few studies on images or other resources. Vuorikari *et al.* (2007) investigated multi-lingual tags in a tagging system on learning resources. They concluded that multi-lingual tags are a kind of resource for language learning, better organizing the multi-lingual resource, and improving user experience. Eleta and Golbeck (2012) studied online art images. Their tags are in English and Spanish. The results suggested that cross-lingual tags help to describe resources from different cultural perspectives (Klavans *et al.*, 2011). They opined that understanding and comparing cross-lingual tags in tagging systems is a prerequisite to improving cultural diversity. Eleta (2011) suggested that cross-lingual tagging hardly decreased the agreement reached within only one language and showed that different types of paintings received different tagging agreements. Jung (2010) applied a framework on cross-lingual tagging analysis and argued that cross-lingual tagging would improve cross-lingual information retrieval. Regarding users' searches, Ruiz and Chin (2010) have found that English tags are more suitable than tags in other languages.

Most existing studies showed low overlap or match between the tags and the subject terms on book annotation (Lee and Schleyer, 2012; Wu *et al.*, 2013). Perhaps social tagging is inherently different from subject indexing, making comparison futile. Social tagging describes different aspects of a resource, as with Dublin Core, but subject headings reveal only the subject of a resource. Social tags of the same resource from different systems seem more comparable. Cross-lingual tags and resources are prevalent on the internet today (Gracia *et al.*, 2012; Jung, 2010; Stiller *et al.*, 2011; Wu *et al.*, 2013). Cross-lingual tags help with cross-lingual retrieval and recommendation, as well as cultural feature revealing (Eleta and Golbeck, 2012; Ruiz and Chin, 2010). Several studies claimed that the tags in different languages could affect search performance (Ruiz and Chin, 2010). The resource category may make a difference too.

Based on the analysis above, this study aims to compare Chinese and English tags of the same books to determine their similarities and differences at a cross-lingual level.

3. Research design

The aim of this study is to estimate the similarity of cross-lingual tags for the same books across social tagging systems. To simplify the research, English tags and Chinese tags are chosen to complete the comparative study. Therefore, three book tagging systems are chosen to collect data: a typical Chinese book tagging system, Douban Reading and two typical English book tagging systems, Amazon and LibraryThing. The main site, Douban.com, is 127th in global rank and 25th in China. Users of Douban are very active. LibraryThing.com is 13,875th in global rank and 5,195th in the USA. Amazon.com is sixth in global rank and fourth in the USA. The ranking information was collected from Alexa (www.alexa.com) on 13 March 2015.

To simplify the description below about cross-lingual tags for one book, this study defines the following concepts:

- *TS* is the abbreviation for a tag set for describing a book in a tagging system;
- *TSC* is the abbreviation for a tag set of a book in a Chinese book tagging system; and
- *TSE* is the abbreviation for a tag set of a book in an English book tagging system.

3.1 Research questions

For a given book, Chinese taggers usually annotate with Chinese tags and English taggers usually annotate with English tags. This study has a clear and specific aim – to determine whether language will affect the book tagging results of a given book. At the TS level, what is the average similarity between the TSCs and TSEs of a given book? Research question one is as follows:

RQ1. What is the average similarity between the TSCs and TSEs of a given book estimated by Jaccard's coefficient at the TS level?

This paper aims to determine whether book category will also affect the degree of similarity. Therefore, research question two is as follows:

RQ2. Do book categories affect the average similarity between the TSCs and TSEs of a given book estimated by Jaccard's coefficient at the TS level?

3.2 Data collection

Data are collected from Douban Reading, LibraryThing and Amazon. Douban is a widespread social annotation system in China; its book section, Douban Reading (http://book.douban.com) – hereinafter referred to as Douban – has abundant book tags and book comments. Douban builds a taxonomy for book classification combining popular tags and manual proofreading. They classify books into six categories: economy and management, science and technology, popular, life, culture and literature. LibraryThing is popular among book lovers worldwide. Nearly 1,700,000 users organize and share books on LibraryThing with social tags. Amazon is famous for its e-commerce; the books listed there gather many related tags and comments, too. Douban provides a maximum of eight top tags per book for users; Amazon provides a maximum of 12, and LibraryThing provides a maximum of 30 default tags.

In total, 1,200 TSCs and 2,400 TSEs for 1,200 books were collected from these three websites. The books selected have both Chinese and English translations and tags that can be found on Douban, Amazon and LibraryThing. Both Amazon and LibraryThing have the same English translation of the selected books. The 1,200 books were evenly chosen from the six categories on Douban.

According to the 1,200 titles and their translated versions, book tags come from Douban, Amazon and LibraryThing. On Douban, the default tags and tagging frequencies (in descending order) of the books were collected. On Amazon, the tag cloud was searched, books were found, screen shots were collected about the tag information for the books, and optical character recognition (OCR) software was used to extract the tags and frequencies (in descending order); the errors caused by the OCR software were manually corrected. On LibraryThing, all the default tags with tagging frequency (not in any particular order) were collected (see the TSEs of the sample in Table I). All 1,200 records are collected between 1 and 15 April 2013. The default tags from the websites were collected to answer the research questions.

3.3 Data analysis

After data collection, the top eight tags are selected as new TSs by their tagging frequency and sorted into descending order from the original ones. A semi-colon (;) is added to separate tags, and all parenthesises are removed. For example, "audiobook (26) business (421)" will be changed to "business; audiobook". Unrecognizable text or meaningless codes, such as @ and +, as well as languages other than Chinese and English, are removed. After that, Chinese tags are translated into English tags using Google Translate (http://translate.google.cn) to transform a cross-lingual problem into a mono-lingual problem and keep the experiment results as original as possible. To avoid errors caused by Google Translate, the translations are checked manually, according to the resource, in Chinese. The Porter stemming algorithm (http://tartarus.org/~martin/PorterStemmer/) is then adopted to further process the data and reduce experiment errors caused by spelling mistakes or tenses (see the sample of final tag data in Table II).

Next, the similarity between the TSEs is analysed. Many studies have compared tags with tags or subject headings (Klavans *et al.*, 2011; Lee and Schleyer, 2012; Lu *et al.*, 2010; Wu *et al.*, 2013). Considering that Jaccard's coefficient is a frequently used method in existing studies, this method is adopted to measure the similarity (or overlap) between the tags of different translations of books. The equation of Jaccard's coefficient is:

Websites	Tag sets	URLs
Douban	{个人管理 – personal management (5,379), 励志–inspirational (3,235), 高效能人士的七个习惯–*The 7 Habits of Highly Effective People* (3,209), 管理–management (2,920), 时间管理–time management (2,438), 效率–efficiency (1444), 成功学–success (1,199), 习惯–habits (1,121)}	http://book.douban.com/subject/1048007/
Amazon	{personal development (200), leadership (160), success (113), time management (107), self-help (107), self improvement (84), inspirational (81), business (66), professional development (53), psychology (34) productivity (19), management (12)}	http://url.cn/dJyC7w[a]
LibraryThing	{audiobook (26), business (421), career (36), character (79), communication (28), ebook (28), effectiveness (39), habits (48), inspiration (29), inspirational (37), leadership (349), management (209), motivation (52), motivational (39), non-fiction (490), organization (73), own (50), personal development (225), personal growth (73), philosophy (29), productivity (137), psychology (213), read (71), reference (45), self improvement (222), self-development (39), self-help (682), success (138), TBR (43), time management (123)}	www.librarything.com/work/3319

Note: [a] This is the short linkage generated from the original URL (www.amazon.com/tag/personal%20development/products/ref=tag_dh_istp)

Table I.
Sample of original tag data from the three websites

Websites	Tags	URLs
Douban	person manag; inspire; the seven habit of highli effect people; manag; time manag; effici; success; habit	http://book.douban.com/subject/1048007/
Amazon	person develop; leadership; success; time manag; self-help; self improve; inspire; busi	http://url.cn/dJyC7w
LibraryThing	self-help; non-fiction; busi; leadership; person develop; self improve; psychology; manag	www.librarything.com/work/3319

Table II.
Sample of the processed tag data from the three websites

$$\text{Jaccard}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \times 100\%, \ (X \cup Y \neq \emptyset) \quad (1)$$

In equation (1), X, Y, respectively, represent a tag set. |X∩Y| represents the amount of the same elements the two sets share; |X∪Y| represents the amount of all the distinct elements that the two sets contain. The quotient is recorded as Jaccard's coefficient. For example, the Jaccard's coefficients of the sample tags shown in Table II between Douban and Amazon are calculated below. The same tags are: "inspir", "time manag" and "success". There are eight tags in Douban for the book and eight in Amazon. J is computed according to equation (1) as:

$$J = \frac{3}{98 + 8 - 30} \times 100\% = 23.1\%$$

In this case, the tagging similarity between Douban and Amazon is 23.1 per cent.

This method examines the similarity between books' TSCs and TSEs. The experiment consists of three groups: two cross-lingual groups – the similarity between Douban and Amazon, and the similarity between Douban and LibraryThing – and a mono-lingual group – the similarity between Amazon and LibraryThing. The results of these three groups will help analyse and answer the first research question. The six categories are taken into consideration to recheck the results. The new findings will help analyse and answer the second research question. The homogeneity test of variance and analysis of variance are conducted on the experimental data using SPSS 20.

4. Findings

Similarity computing provides three groups of similarity between tags from different groups: Douban versus LibraryThing, cross-lingual; Douban versus Amazon, cross-lingual; and LibraryThing verses Amazon, mono-lingual. Then, a statistical analysis unfolds in two aspects: the general level and the category level. The general level analysis answers *RQ1* and the category level answers *RQ2*.

4.1 General statistics of tagging similarity

The similarity between the book TSs of the three groups is computed. The homogeneity test of variance and analysis of variance on the three groups of Jaccard's coefficients using SPSS 20 are in Tables III and IV. The Levene statistic is 133.398 and *F*-value = 53.883. The three groups of data are significantly different.

The interval statistics of the similarity in the three groups are in Table V. More than 80 per cent of Jaccard's coefficient values fall into [0, 30.0 per cent], while the rest make up only about 20 per cent. The average values of similarity (or overlap) in the groups are much larger than those estimated in other studies (Lee and Schleyer, 2012; Wu *et al.*, 2013). An explanation is that tags resemble each other more than they resemble controlled vocabularies. But the data also suggest that a small

Levene statistic	df1	df2	Sig.
133.398	2	3,597	0.000

Table III. Homogeneity test of variance

number of books share totally different tags across the three websites, especially between Douban and Amazon. The situation changes a little when it comes to different groups. In Douban–LibraryThing and Amazon–LibraryThing, most of the Jaccard's coefficient values (31.0 and 24.3 per cent) fall in the interval between 10 and 20 per cent; in Douban–Amazon, most of the Jaccard's coefficient values (30.0 per cent) fall between 0 and 10 per cent. Amazon–LibraryThing shows a slight advantage in average similarity. More data are distributed in the high-valued intervals and more evenly. Tags for the same books on LibraryThing and Amazon are more similar to each other.

Table IV.
Single factor analysis of variance

Types	Sum of squares	df	Mean squares	*F*	Significance
Between groups	0.663	2	0.332	53.883	0.000
Within group	22.142	3,597	0.006		
Total	22.805	3,599			

Table V.
Interval statistics of the similarity in the three groups

Groups	Interval	Average (%)	Fre.
Douban–Amazon	0-0.1	4.3	*574 (219)****
	0.1-0.2	14.8	340
	0.2-0.3	24.3	177
	0.3-0.4	33.8	86
	0.4-0.5	44.6	16
	0.5-0.6	51.1	5
	0.6-0.7	63.3	2
Overall in Douban–Amazon	0-0.8	*13.2****	1,200
Douban–LibraryThing	0-0.1	5.9	266 (40)
	0.1-0.2	14.8	*374*
	0.2-0.3	24.2	327
	0.3-0.4	34.2	168
	0.4-0.5	43.7	43
	0.5-0.6	50.6	18
	0.6-0.7	60.0	4
Overall in Douban–LibraryThing	0-0.8	*19.8****	1,200
Amazon–LibraryThing	0-0.1	3.8	280
	0.1-0.2	14.2	*292 (129)****
	0.2-0.3	23.2	291
	0.3-0.4	33.4	221
	0.4-0.5	45.3	97
	0.5-0.6	52.8	2
	0.6-0.7	60.0	16
	0.7-0.8	77.8	1
Overall in Amazon–LibraryThing	0-0.8	*20.7****	1,200
Total	0-0.8	17.9	3,600

Note: The numbers in parentheses represent how many of the Jaccard's coefficients are equal to 0 in each [0, 0.1]; *** represents $p < 0.001$

The descriptive statistics of the three groups' Jaccard's coefficients are in Table VI. All the Jaccard's coefficient values are in [0, 80.0 per cent]. The highest Jaccard's coefficient value is 78 per cent and the lowest is 0 among the three groups. Amazon–LibraryThing gets more none-zero Jaccard's coefficients, and the data are distributed more evenly. Amazon–LibraryThing and Douban–LibraryThing have the highest average similarity (more than 20 per cent). Amazon-Douban has the lowest (13.0 per cent).

To sum up, Douban–Amazon gains the lowest similarity level of the three groups; Amazon–LibraryThing gains the highest, closely followed by Douban–Amazon. The above findings suggest that cross-lingual tagging systems can get high similarity in tagging comparison.

4.2 Similarity between English and Chinese tags at the category level

The taxonomy of books in Douban allows for an in-depth investigation of the differences caused by book categories. The average Jaccard's coefficient values in each category are calculated in each of the three groups. The results of the homogeneity test of variance and analysis of variance are shown in Tables VII and VIII. The test results show an F-value of Levene = 15.659; all F-values in between-subject effects are more than 4, suggesting that the three groups of similarity between six categories are significantly different, too.

Table IX shows that, within groups, different categories differ in average similarity. Two categories, economics and management and science and technology, show the highest average similarity and standard deviations among all three groups, while the other categories suggest low average similarity and standard deviations. The mono-lingual group, Amazon–LibraryThing, gains the highest average similarity and standard deviations in almost all the categories. Douban–Amazon shows the lowest similarity in all categories.

5. Discussion

This study compared the tagging of 1,200 books in three groups – Douban–Amazon, Douban–LibraryThing and Amazon–LibraryThing – to investigate book tagging data from a cross-lingual perspective. This section reviews the results and identifies some insights and implications of this study.

Descriptive statistics	Douban–Amazon	Douban–LibraryThing	Amazon–LibraryThing
Max (%)	67	60	78
Min (%)	0	0	0
Median (%)	11	17	23
Average (%)	13	20	21
SD	0.11	0.11	0.14

Table VI. Descriptive statistics of Jaccard's coefficients of each group

F	df1	df2	Significance
15.659	17	3,582	0.000

Notes: [a] Design: intercept + group + category + group × category

Table VII. Levene's test of equality of error variances[a]

Table VIII.
Tests of between-subject effects

Source	Type III sum of squares	df	Mean Square	F	Significance
Corrected model	11.570[a]	17	0.681	55.151	0.000
Intercept	115.514	1	115.514	9,360.566	0.000
Group	4.071	2	2.035	164.927	0.000
Category	6.991	5	1.398	113.297	0.000
Group × Category	0.509	10	0.051	4.123	0.000
Error	44.204	3,582	0.012		
Total	171.287	3,600			
Corrected total	55.774	3,599			

Note: [a] $R^2 = 0.190$ (Adjusted $R^2 = 0.186$)

Table IX.
Mean Jaccard's coefficients ± STDEV of each category in each group

Categories	Douban–Amazon	Douban–LibraryThing	Amazon–LibraryThing
Economics and management (%)	*19**** ± 11	*23**** ± 13	*25**** ± 11
Science and technology (%)	*21**** ± 11	*26**** ± 13	*28**** ± 14
Popular (%)	9 ± 8	17 ± 8	19 ± 12
Life (%)	9 ± 8	15 ± 9	20 ± 13
Culture (%)	13 ± 10	21 ± 11	20 ± 13
Literature (%)	8 ± 9	16 ± 8	14 ± 14

Note: *** Represents $p < 0.001$

5.1 High similarity can exist between the Chinese book tagging systems and English book tagging systems for a given book

The results suggest that the similarity between mono-lingual tags for a given book is slightly higher than that between cross-lingual tags for the same book. But the cross-lingual group, Douban–LibraryThing, also got quite high similarity simultaneously.

In the cross-lingual groups, Douban–Amazon and Douban–LibraryThing, the average similarity estimated by Jaccard's coefficient differs (13 and 20 per cent, respectively). The Douban–LibraryThing group has achieved nearly the same similarity as the Amazon–LibraryThing group, the mono-lingual group. The comparison suggests that Douban–LibraryThing has a higher level of tag similarity than Douban–Amazon. Between Douban and Amazon, more than one-sixth (219/1,200) of the books have not shared the same tags; between Amazon and LibraryThing the rate is 129/1,200; however, between Douban and LibraryThing the rate is 40/1,200.

The language may not cause tagging differences, but the users will. In Amazon, most of the tagging users are book consumers. When they buy books, they tag them. Users come from different classes of society and may have different tagging habits. On Douban and LibraryThing, most of the users are librarians and readers. They have access to many books. They may have similar tagging habits and behaviours.

5.2 Book category may affect the average similarity between the Chinese book tagging systems and English book tagging systems

The results suggest that the similarity between tags of books written for specialties is higher than that of books written for the general public. The books in economics and management and science and technology share much higher similarity in all three groups. Books in other categories, such as literature and life, show lower similarity.

Perhaps this is because the books of these two categories are for specialists and the tags for the books share like terms: for example, "C++", "black hole" and "supply chain". These tags will improve the average similarity in two ways: specialists are more likely to annotate with similar tags because they know the terminology; these terms decrease the occurrence of synonyms. The tags in other categories are more general: for example, "novel", "fiction" and "thriller". These tags will impair the average similarity in two ways: users from different backgrounds can tag the books, and the tags with several synonyms increase the probability that users will introduce even more synonyms, which cannot be assessed in this study.

To realize the cross-lingual analysis in this study, all the Chinese tags are translated to English tags by Google Translate. To avoid errors caused by machine translation, the translated tags are manually checked before comparison. Even though a bias still exists, considering that the two cross-lingual groups share the same translated Chinese tag data, comparing the results of the two groups can minimize errors and preserve the utility of the findings.

5.3 Insights from the results

Users who have more in common annotate with similar tags. The similarity of the three groups suggests that the tags of different languages can be nearly the same as those of the monolingual group. This means that, while language may affect tagging behaviour, users affect it more. Similarly, book categories cause differences in average similarity among the three groups. The books written for specialists helped the Douban–Amazon group achieve high tagging similarity. At the cross-lingual level, most of the users on Douban and LibraryThing are book lovers and librarians, while Amazon users are always book consumers. The users on Douban and LibraryThing share more common space, such as linguistic habits, than users between Douban and Amazon. At the book category level, those engaged with similar professions tend to share more common knowledge and terminology; they have more in common. Among the six categories, economics and management and science and technology are the more professional subjects. For example, users reading books on economics and management are more likely to have similar backgrounds; they have something in common. Therefore, the tagging similarity is higher, regardless of the tagging language or tagging system type.

5.4 Implications

- *Suggestions for cross-lingual book tagging system design.* Users shape the tagging system the researchers are to design. The designer should consider future users. The links between tags of the same meaning should be well-established to improve tag retrieval. Books written for specialists may have well-established links between different language tags of the same meaning. Users can tag a book without using their native language. The number of tags for each book can be well controlled. For books written for the general public, the system may allow users to annotate with more tags to better interpret the books. Due to the heterogeneity of users, establishing links between different language tags of the same meanings becomes a challenge. For these kinds of books, the systems must grant users more permissions. Determining how best to help users annotate with the same tags (or in other languages) may be one way to control the number of book tags.

- *Tricks for cross-lingual book searching.* When in a cross-lingual environment, users may want to find a book using their own language. They may use more terms to search professional books, such as "C++" or "supply chain"; they may use more general tags and their synonyms to search ordinary books, and make more attempts.
- *Book recommendations.* When it comes to book recommendations, the developers of tagging systems also need to consider book category. The tagging systems should provide more default tags on ordinary books; some of the tags can be synonyms or abbreviations to help users find the books they want.

The aim of this study is to investigate whether English and Chinese tags for the same books share similar meanings from a cross-lingual perspective. It provides a systematic process for cross-lingual analysis of the similarity between a given book's English and Chinese tags. It will help us better understand the relationships between tags in different languages from the same resource. The results of this study can be referred to examine user behaviours and preferences for different systems, taking into consideration individual culture backgrounds in a cross-lingual way. The results and conclusion of this study will benefit cross-lingual information retrieval and book recommendations in tagging systems.

6. Conclusion

With the development of social media, social tagging is gaining popularity among users all over the world. To analyse if English and Chinese tags for the same books share similar meanings, social tags of 1,200 books were collected from one Chinese and two English bookmarking systems. Chinese tags were translated into English using Google Translate. All tags were stemmed via the Porter Stemming Tool. Jaccard's coefficient was adopted to examine the similarity between tags for the same book in different systems. The investigators found that:

- the similarity between mono-lingual tags for a given book is higher than those of cross-lingual tags from different systems; and
- considering category, the similarity between tags of books written for specialties is higher than that of books written for the general public.

To sum up, those who have something in common annotate with tags that are similar. The similarity between the users in tagging systems decides the similarity of the tag sets.

The significance of this study is threefold. First, the comparison of cross-lingual tags for the same resource will benefit multilingual retrieval and book recommendations by library systems and e-commerce corporations with folksonomy. Second, this study will fill in some blanks in cross-lingual tag comparison and tag comparative studies between different tagging systems. Third, the results of this study can be used to examine different user behaviours and preferences for different systems, and their cultural backgrounds, in a cross-lingual way. In addition, its methodology can be applied to the tag comparison between any other two languages.

However, this study has some limitations. To investigate the similarity between TSCs and TSEs of a given book, Google Translate was used to translate Chinese tags to English. The translating process may cause errors despite being manually checked afterwards. For example, in British English, the equivalent of "小说" is "fiction", while in American English it is "novel". All three phrases have the same meaning, but the similarity could not be captured by using Jaccard's coefficient. Also, manual translation according to the English tags can cause bias. Both methods seem insufficient, and the synonyms in mono-lingual tags could not be assessed by Jaccard's coefficient. That affected the similarity between tags in Amazon and LibraryThing, such as the similarity between "novel" and "thriller". One possible solution to these two limitations would be to use a more advanced corpus, such as Wikipedia or WordNet, in future studies. Another possible limitation may be that two translations of a book (English and Chinese) were chosen without regard to its original language, even though most of the 1,200 books were originally published in English. The selection of book translations may slightly affect the precision of this study.

Future work will focus on applying advanced semantic tools like WordNet to the analysis of the relationships to mine further information, including syntax information. Yet another mountain to climb will be determining how to manage cross-lingual books and their tagging data from different districts, and better serve users worldwide.

References

Ames, M. and Naaman, M. (2007), "Why we tag: motivations for annotation in mobile and online media", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 971-980.

Bartley, P. (2009), "Book tagging on LibraryThing: how, why, and what are in the tags?", *Proceedings of the American Society for Information Science and Technology*, Vol. 46 No. 1, pp. 1-22.

Carman, M.J., Baillie, M. and Crestani, F. (2008), "Tag data and personalized information retrieval", *Proceedings of the 2008 ACM Workshop on Search in Social Media*, ACM, New York, NY, pp. 27-34.

Catarino, M.E. and Baptista, A.A. (2010), "Relating folksonomies with Dublin Core", *International Journal of Metadata, Semantics and Ontologies*, Vol. 5 No. 4, pp. 285-295.

Chen, Y.N. and Ke, H.R. (2013), "An analysis of users' behaviour patterns in the organisation of information: a case study of CiteULike", *Online Information Review*, Vol. 37 No. 4, pp. 638-656.

De Meo, P., Ferrara, E., Abel, F., Aroyo, L. and Houben, G.-J. (2013), "Analyzing user behaviour across social sharing environments", *ACM Transactions on Intelligent Systems and Technology*, Vol. 5 No. 1, p. 14.

DeZelar-Tiedman, C. (2011), "Exploring user-contributed metadata's potential to enhance access to literary works", *Library Resources & Technical Services*, Vol. 55 No. 4, pp. 221-233.

Eleta, I. (2011), "Art images and multilingual social tagging: a museum without borders", available at: http://drum.lib.umd.edu/bitstream/1903/11395/1/LAMP-TR-156 pdf (accessed 12 May 2015).

Eleta, I. and Golbeck, J. (2012), "A study of multilingual social tagging of art images: cultural bridges and diversity", *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, ACM, New York, NY, pp. 695-704.

Elhussein, M. and Nakata, K. (2010), “Cross-lingual information retrieval as a side effect of tagging”, *2010 IEEE Second International Conference on Social Computing (SocialCom)*, IEEE, Piscataway, NJ, pp. 582-586.

Elhussein, M. and Nakata, K. (2012), “Analysing the factors that influence tag choice based on semiotic analysis and activity theory”, *2012 International Conference on Social Informatics (SocialInformatics 2012)*, IEEE, Piscataway, NJ, pp. 96-105.

Farooq, U., Kannampallil, T.G., Song, Y., Ganoe, C.H., Carroll, J.M. and Giles, L. (2007), “Evaluating tagging behaviour in social bookmarking systems: metrics and design heuristics”, *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, ACM, New York, NY, pp. 351-360.

Gelernter, J. (2007), “A quantitative analysis of collaborative tags: evaluation for information retrieval – a preliminary study”, *2007 International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2007)*, IEEE, Piscataway, NJ, pp. 376-381.

Golbeck, J., Koepfler, J. and Emmerling, B. (2011), “An experimental study of social tagging behaviour and image content”, *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 9, pp. 1750-1760.

Golder, S.A. and Huberman, B.A. (2006), “Usage patterns of collaborative tagging systems”, *Journal of Information Science*, Vol. 32 No. 2, pp. 198-208.

Golub, K., Lykke, M. and Tudhope, D. (2014), “Enhancing social tagging with automated keywords from the dewey decimal classification”, *Journal of Documentation*, Vol. 70 No. 5, pp. 801-828.

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P. and McCrae, J. (2012), “Challenges for the multilingual web of data”, *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 11, pp. 63-71.

Guerra, R. and LaPlante, R. (2011), “Art images online: leveraging social tagging and language for browsing”, available at: www.umiacs.umd.edu/research/t3/documents/UMd-T3-CNI-December-2011 pdf (accessed 12 May 2015).

Gupta, M., Li, R., Yin, Z. and Han, J. (2010), “Survey on social tagging techniques”, *ACM SIGKDD Explorations Newsletter*, Vol. 12 No. 1, pp. 58-72.

Heckner, M., Mühlbacher, S. and Wolff, C. (2008), “Tagging tagging: analysing user keywords in scientific bibliography management systems”, *Journal of Digital Information*, Vol. 9 No. 2, pp. 1-19.

Jung, J.J. (2010), “Matching multilingual tags based on community of lingual practice from multiple folksonomy: a preliminary result”, in García-Pedrajas, N., Herrera, F., Fyfe, C., Benítez, J.M. and Ali, M. (Eds), *Trends in Applied Intelligent Systems*, Springer-Verlag, Berlin/Heidelberg, pp. 39-46.

Jung, J.J. (2012), “Discovering community of lingual practice for matching multilingual tags from folksonomies”, *The Computer Journal*, Vol. 55 No. 3, pp. 337-346.

Kakali, C. and Papatheodorou, C. (2010), “Exploitation of folksonomies in subject analysis”, *Library & Information Science Research*, Vol. 32 No. 3, pp. 192-202.

Ke, H.R. and Chen, Y.N. (2012), “Structure and pattern of social tags for keyword selection behaviours”, *Scientometrics*, Vol. 92 No. 1, pp. 43-62.

Kipp, M.E. and Campbell, D.G. (2006), “Patterns and inconsistencies in collaborative tagging systems: an examination of tagging practices”, *Proceedings of the American Society for Information Science and Technology*, Vol. 43 No. 1, pp. 1-18.

Klavans, J.L., Guerra, R., LaPlante, R., Stein, R. and Bachta, E. (2011), "Taming social tags: computational linguistic analysis of tags for images in museums", available at: http://drum.lib.umd.edu/bitstream/1903/11394/1/LAMP-TR-155 pdf (accessed 12 May 2015).

Lawson, K.G. (2009), "Mining social tagging data for enhanced subject access for readers and researchers", *Journal of Academic Librarianship*, Vol. 35 No. 6, pp. 574-582.

Lee, D.H. and Schleyer, T. (2012), "Social tagging is no substitute for controlled indexing: a comparison of medical subject headings and cite U like tags assigned to 231,388 papers", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 9, pp. 1747-1757.

Li, X., Snoek, C.G. and Worring, M. (2008), "Learning tag relevance by neighbor voting for social image retrieval", *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, ACM, New York, NY, pp. 180-187.

Lin, C.S. and Chen, Y.F. (2012), "Examining social tagging behaviour and the construction of an online folk sonomy from the perspectives of cultural capital and social capital", *Journal of Information Science*, Vol. 38 No. 6, pp. 540-557.

Lu, C., Park, J. and Hu, X. (2010), "User tags versus expert-assigned subject terms: a comparison of Library Thing tags and library of congress subject headings", *Journal of Information Science*, Vol. 36 No. 6, pp. 763-779.

Petek, M. (2012), "Comparing user-generated and librarian-generated metadata on digital images", *OCLC Systems & Services*, Vol. 28 No. 2, pp. 101-111.

Peters, I. (2009), *Folksonomies: Indexing and Retrieval in Web 2.0*, De Gruyter Saur, Berlin.

Rolla, P.J. (2009), "User tags versus subject headings", *Library Resources & Technical Services*, Vol. 53 No. 3, pp. 174-184.

Rorissa, A. (2010), "A comparative study of Flickr tags and index terms in a general image collection", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 11, pp. 2230-2242.

Ruiz, M.E. and Chin, P. (2010), "Users' seeking behaviour and multilingual image tags", *Proceedings of the American Society for Information Science and Technology*, Vol. 47 No. 1, pp. 1-2.

Santos-Neto, E., Condon, D., Andrade, N., Iamnitchi, A. and Ripeanu, M. (2009), "Individual and social behaviour in tagging systems", *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, ACM, New York, NY, pp. 183-192.

Spiteri, L.F. (2013), "The structure and form of folksonomy tags: the road to the public library catalog", *Information Technology and Libraries*, Vol. 26 No. 3, pp. 13-25.

Stiller, J., Gäde, M. and Petras, V. (2011), "Is tagging multilingual? A case study with BibSonomy", *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, ACM, New York, NY, pp. 421-422.

Strohmaier, M., Körner, C. and Kern, R. (2010), "Why do users tag? Detecting users' motivation for tagging in social tagging systems", *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, AAAI, Palo Alto, CA, pp. 339-342.

Strohmaier, M., Körner, C. and Kern, R. (2012), "Understanding why users tag: a survey of tagging motivation literature and results from an empirical study", *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 17, pp. 1-11.

Tennis, J.T. (2006), "Social tagging and the next steps for indexing", *Advances in Classification Research Online*, Vol. 17 No. 1, pp. 1-15.

Thomas, M., Caudle, D.M. and Schmitz, C.M. (2009), "To tag or not to tag?", *Library Hi Tech*, Vol. 27 No. 3, pp. 411-434.

Vuorikari, R., Ochoa, X. and Duval, E. (2007), "Analysis of user behaviour on multilingual tagging of learning resources", *Proceedings of the 1st Workshop on Social Information Retrieval for Technology-Enhanced Learning & Exchange*, pp. 6-17.

Wan, Y., Yang, X., Liu, Z., Ma, J., Li, X., Ouyang, C., Yu, Y. (2013), "Study on user behaviour and social tagging", *Proceedings of the International Conference on Computer, Networks and Communication Engineering (ICCNCE 2013)*, Vol. 30, pp. 1-4.

Wetzker, R., Zimmermann, C. and Bauckhage, C. (2008), "Analyzing social bookmarking systems: a del.icio.us cookbook", in *Proceedings of the ECAI 2008 Mining Social Data Workshop*, pp. 26-30.

Wu, D., He, D., Qiu, J., Lin, R. and Liu, Y. (2013), "Comparing social tags with subject headings on annotating books: a study comparing the information science domain in English and Chinese", *Journal of Information Science*, Vol. 39 No. 2, pp. 169-187.

Yi, K. and Chan, L.M. (2009), "Linking folksonomy to Library of Congress subject headings: an exploratory study", *Journal of Documentation*, Vol. 65 No. 6, pp. 872-900.

Corresponding author
Chengzhi Zhang can be contacted at: zhangcz@njust.edu.cn